CASCADE CUP ROUND 3 - EXPLORATORY DATA ANALYSIS
Team: Tony-buddies

Introduction to the Problem statement:

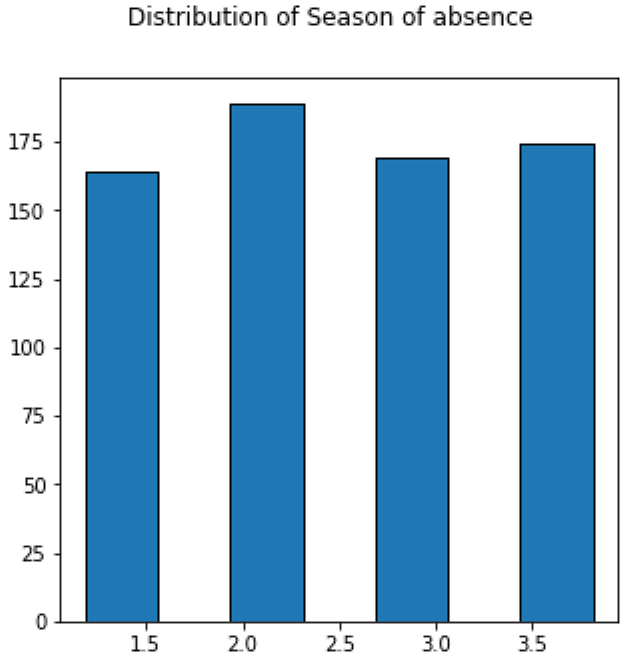
- ➢ Absenteeism is the pattern of unplanned absences from work, and we analyze the 28 reasons employees gave for absences at a courier company in Brazil.
- ➢ The reason for absences includes 21 ICD(International Code for Diseases) reasons like Diseases of the eye, ear, nervous system, etc., and 7 non-ICD reasons like physiotherapy, a medical consultation, and unjustified absence, among others.
- ➢ For each employee, we have their ID number, the number of hours of absentee time, and many other qualities of the individual like the education, number of children, age, distance of residence, etc.

Initial data cleaning:

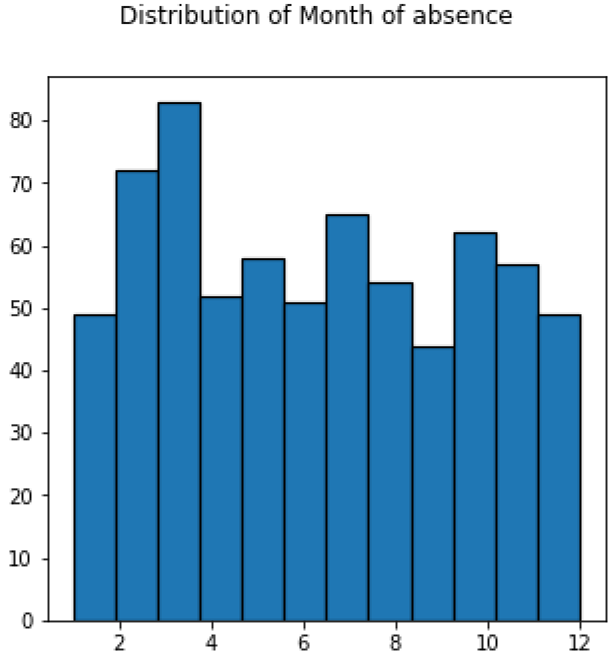
- ➢ We see that there are no null values in the data, but there were a few unexpected values.
- ➢ There were 3 rows with the month as zero, but the range for the month was 1 to 12. So we ignored those rows as corrupted data. All three of these months have zero absenteeism as well, so we can safely ignore these rows as not useful.
- ➢ We have 28 reasons for absenteeism described, but 40 rows have a reason as 0,
- ➢ And all 40 of them have zero absentee time as well, so we are ignoring these rows.
- ➢ We are trying to analyze reasons for absenteeism, but 1 row still has zero absenteeism time, which is not useful, so we ignore that row.
- ➢ After these changes, we are left with 696 rows.

Analyzing employee distribution:
Temporal distribution of data:

- ➢ The seasons of absenteeism are fairly spread out in our dataset. This will allow for a fair analysis of absenteeism and the reasons throughout the year.

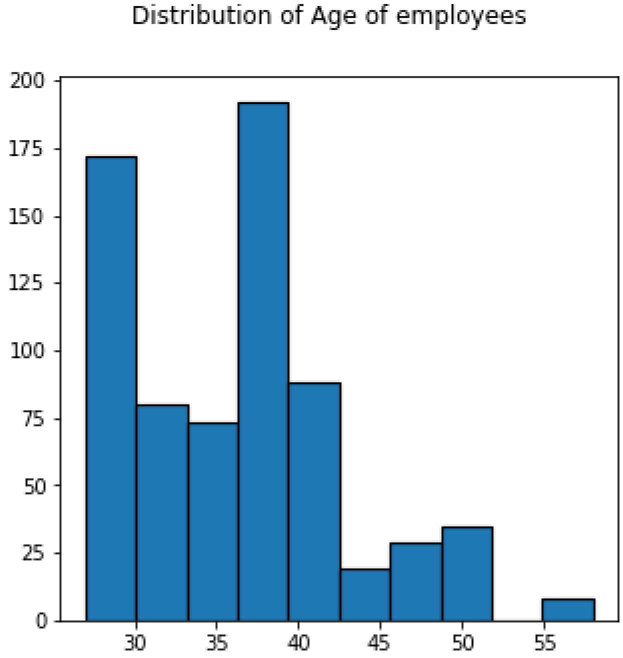


➢ This is also clarified by observing the distribution of our data across all the months is fairly uniform. This means that there is **no temporal bias** in our data.



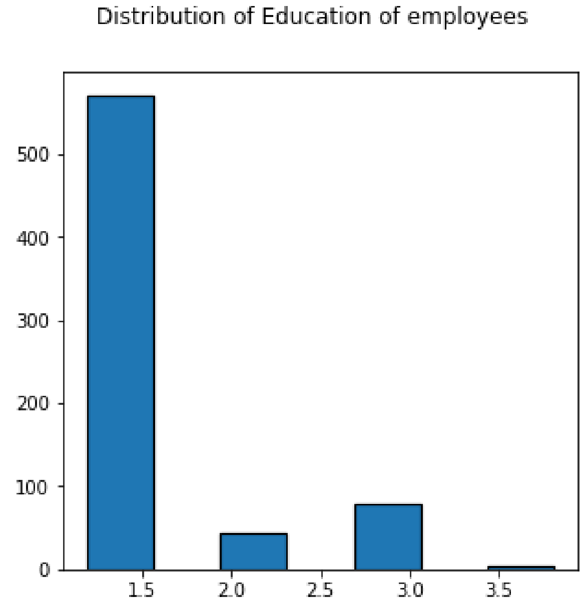
Distribution Age of employees

➢ The age of employees in our data is mostly concentrated around the late 20s to early 40s. The data is spare after the age of 42. Hence we interpret this data as the absenteeism analysis for the average middle-aged employees of the company.



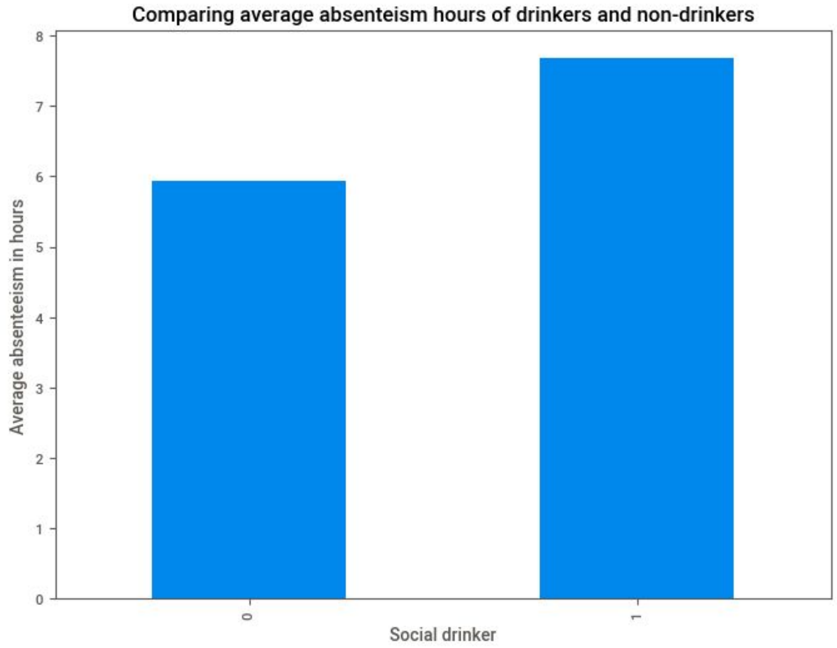
Education of employees:

- We observe that the majority of the employees are educated only up to high school. This may be due to the type of company selected for analysis.
- For reference, the education value of 1 is for High school, 2 is for graduate, 3 is for post-graduate, and 4 is for masters and doctors.



Distribution of Social Smokers and Social drinkers and the effect on absenteeism:

- 93% of the employees are non-smokers, showing the bias in our data towards non-smokers. Thus the analysis based on smoking is not useful for us.
- 57% of the employees are social drinkers, and 43% are non-drinkers. This shows the data is evenly distributed for us to compare the absenteeism of drinkers and non-drinkers.
- We can see that social drinkers, on average, have more absenteeism hours (8 hours) as compared to non-drinkers (6 hours). This shows the negative impact of drinking on their work ethics and job performance.

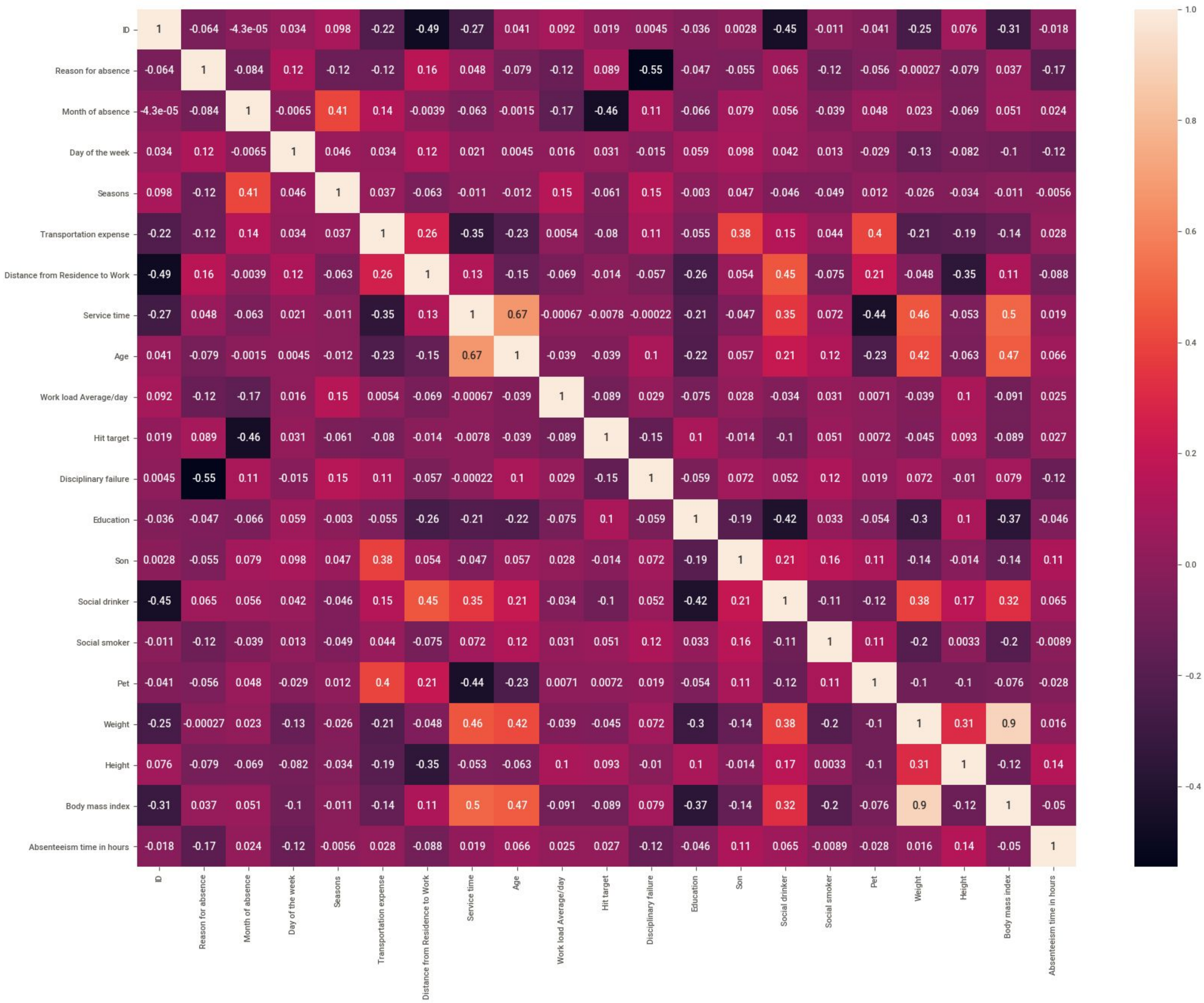


Correlation of employee features:

We plotted a correlation map, and some of the features with the highest correlation are:

- Age and service time
- Age and Body mass index
- Social drinker and distance from residence to work
- Weight and service time
- Disciplinary failure and reason for absence
- Social drinker and education
- Hit target and month of absence
- Seasons and month of absence

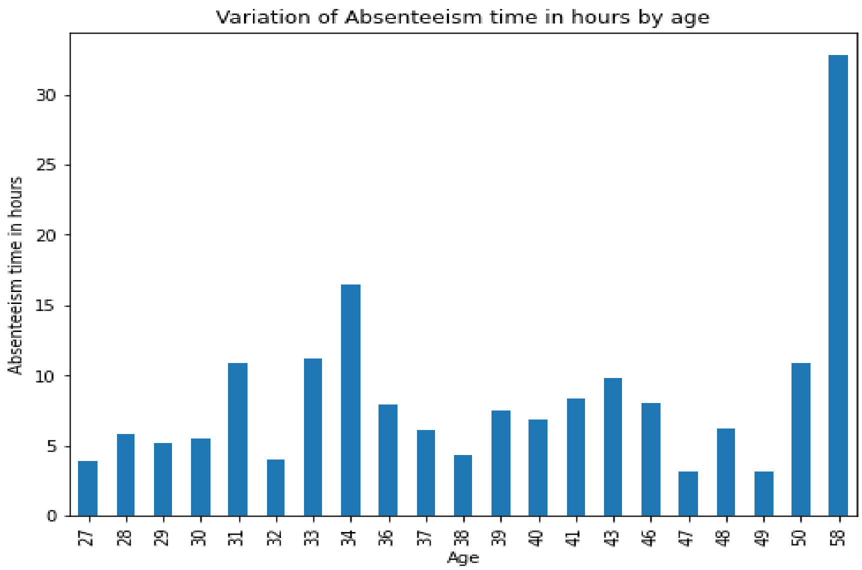
As we can see, absentee time is not strongly correlated with any particular feature of the employee, but let’s try to understand how absentee time is depending on the combination of all the features given.



Variation of Absentee time with the age of an employee:

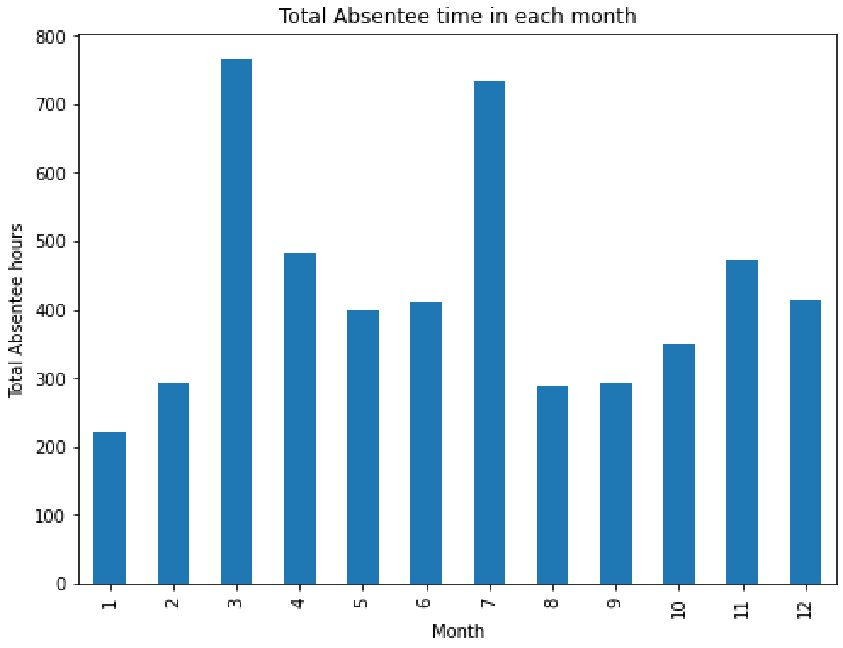
➢ We see that, though most of the employees are middle-aged, the average absentee time is the highest for the older employees of age 58.

➢ Other than that anomaly, the absentee time is uniformly distributed across all ages, indicating no bias in our data for any particular age group.



Variation of Absenteeism across different months of the year:

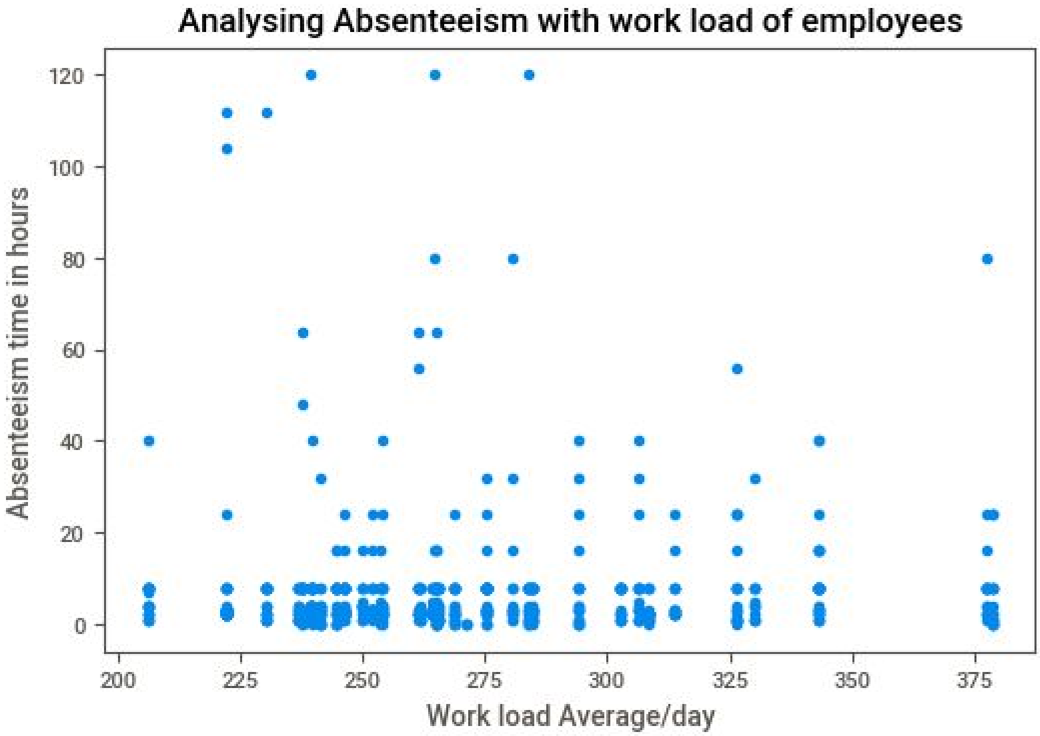
- We observe that the total absentee time in the months of March(3) and July(7) are the highest by a considerable margin.
- In the month of March, there were 797 hours of absentee time. The reasons for this are mainly injuries/poisoning due to external causes, diseases related to muscles, and skin. This indicates that the high absenteeism in March might be due to climatic factors.
- In July, there were 750 hours of absentee time. The primary reasons for absenteeism here are diseases of the circulatory and nervous systems.



Variation of Absenteeism with an average workload of employees:

- We observe that the Absenteeism hours for employees doesn't depend on the workload like one would expect.

- ➢ This insight is from the plot of absentee time with the workload of employees. This plot is relatively uniform, barring a few anomalies.
- ➢ This indicates that absenteeism is not merely a product of the workload or working conditions. It might be due to other personal/external factors affecting the employees.



Conclusion:

- ➢ From the above analysis, we see that the absenteeism and the reasons for absenteeism seem to be related to particular employees' features. Beyond that, they depend on the external environmental factors affecting them. The absenteeism is observed to not depend on the workload, meaning that the employees seem reasonably satisfied with the company.
- ➢ The employees (drinking/smoking) have been shown to affect their work performance negatively. We see that employees' habit of social drinking had adverse effects on that employee's Absenteeism.
- ➢ We also observe that the absentee time in March and July is high, and the reasons tend to be external factors affecting employees' health. The given data doesn't have any temporal bias, which provides this insight with a little more confidence.