Analysis of Absenteeism

Cascade Cup’20

Abstract

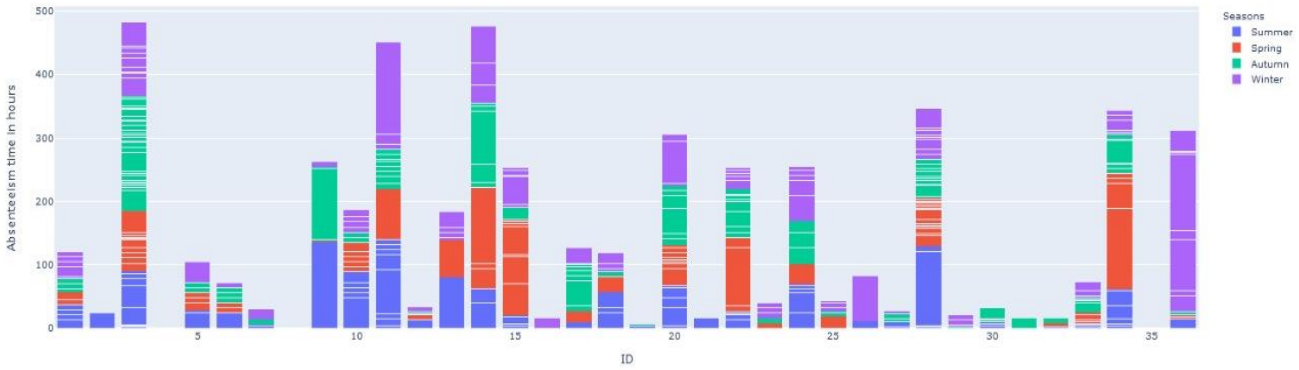
Absenteeism is a habitual pattern of absence from a duty or obligation without good reason. Generally, absenteeism is unplanned absences. If a workplace exhibits high degree of absenteeism there is a problem. It has been viewed as an indicator of poor individual performance, as well as a breach of an implicit contract between employee and employer.

We have done an EDA of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil.

Team: DS_Wizards [Anish S Ghiya & Aadarsh Sudhir Ghiya]

Contents

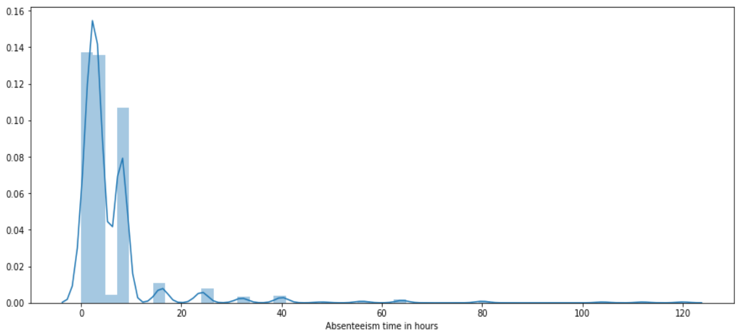
Distribution of hours throughout the dataset by Employee

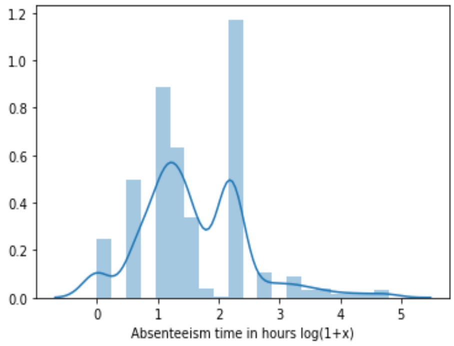


As seen in the graph, Employee with ID : 4, 8 and 35 have taken no leaves in the 3 years. Employee with ID 3 and 14 have the highest absenteeism hours. Employee 16, 36 takes his maximum leave during winter. Employee 2 takes his leaves in Summer. More can be inferred using the graph above.

Distribution of hours throughout the dataset

The original distribution of the target (absenteeism in hours) is not a normal distribution and a heave positive skew was detected which may cause problems in the machine learning models. A log(1+x) transformation helped attain a more normal distribution.

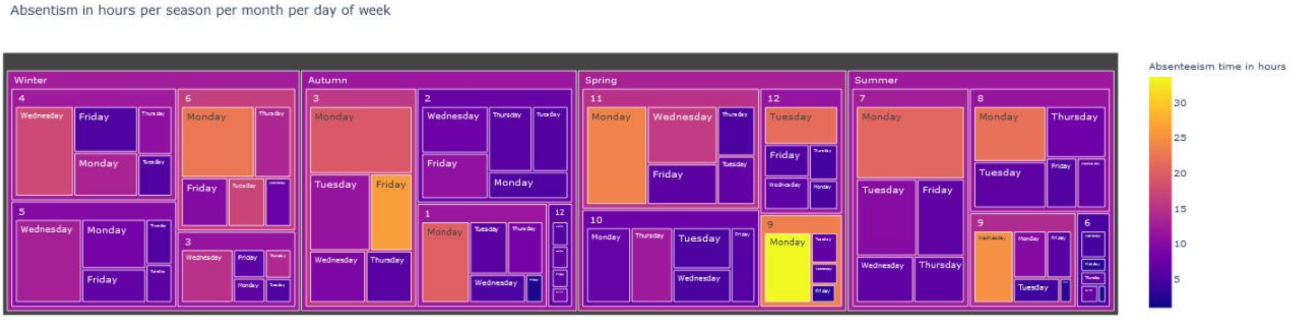




Distribution based on Season, Months and Day of Week

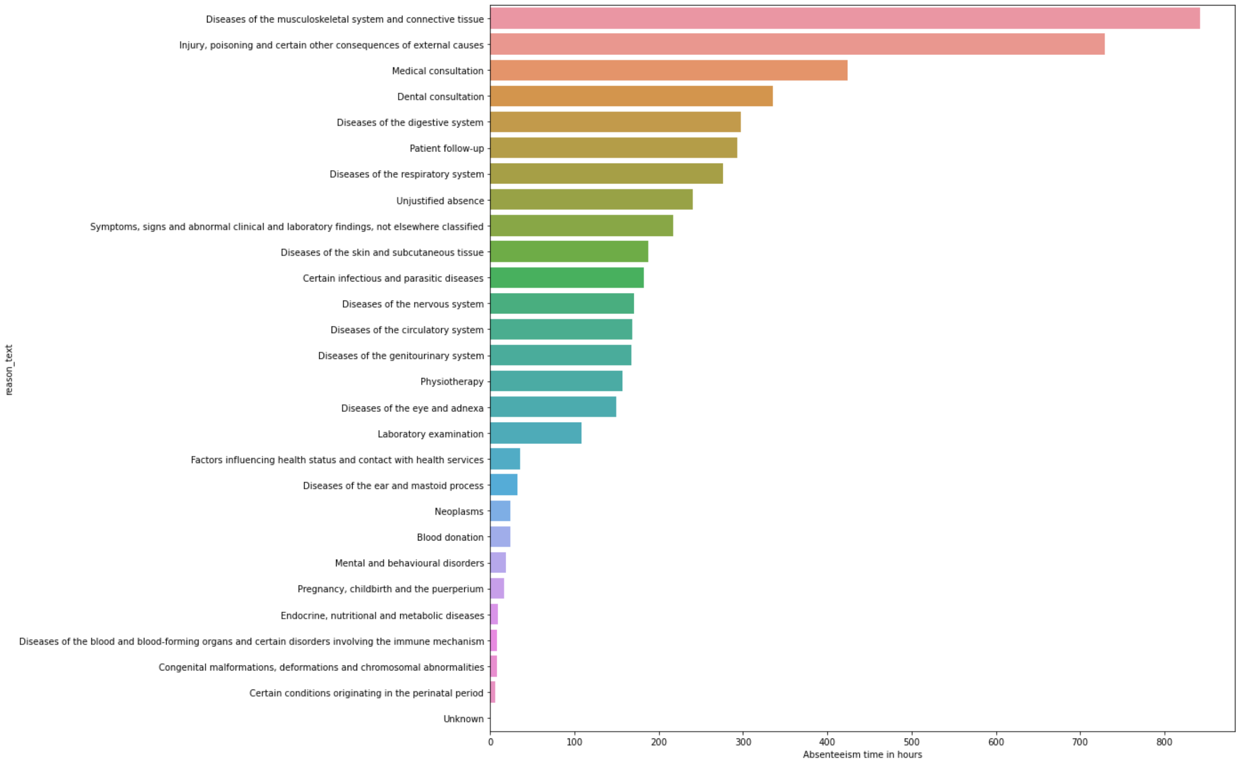
For this we plotted a tree map to see the distribution. Based on the plot we noticed that Winter has the maximum hours of absenteeism.

In Winter the 4th month, has the highest absenteeism hours. During Autumn the 3rd month, has the highest number of hours in absenteeism. A point to note is that Month 3 comes in both Autum and Winter.

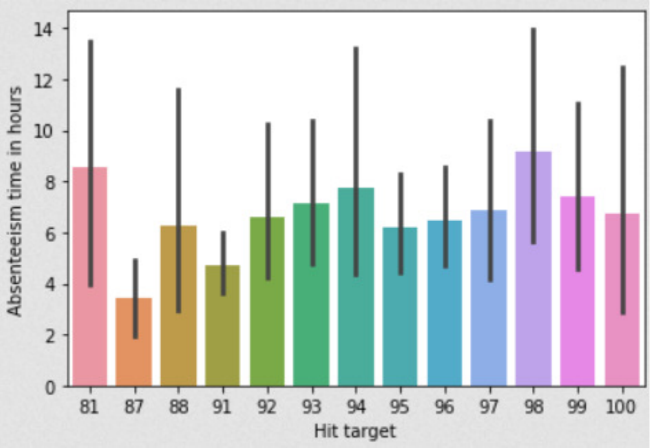


Distribution based on International Code of Diseases (ICD)

Disease 13. Diseases of the musculoskeletal system and connective tissue and 19. Injury, poisoning and certain other consequences of external causes are the reason for maximum absenteeism.

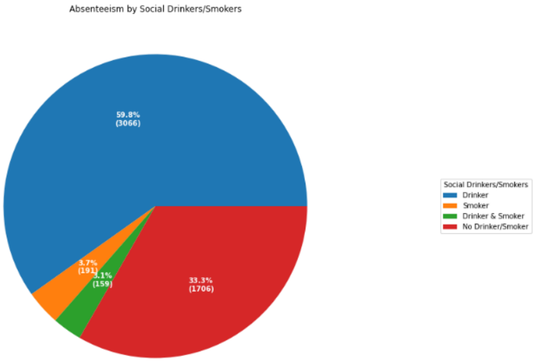


Distribution of Hit target based absentism



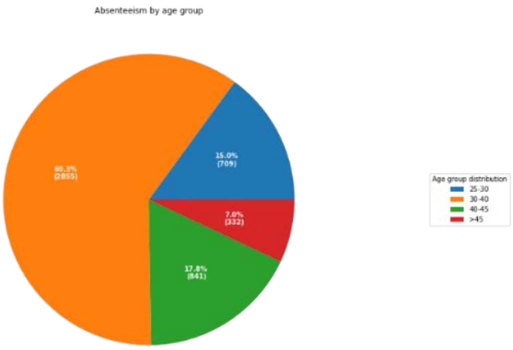
Percentage distribution based social drinkers and smokers

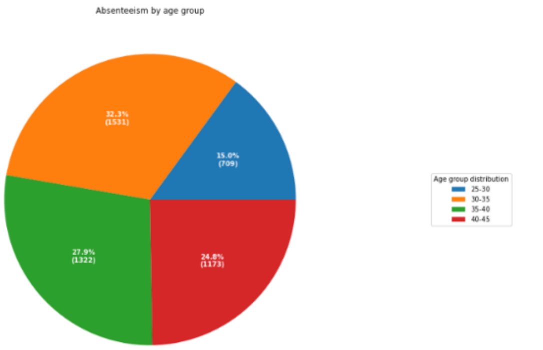
More than 50% of the absenteeism is because of social drinkers, and also non drinkers and non smokers make up 33.33% of the absenteeism hours.



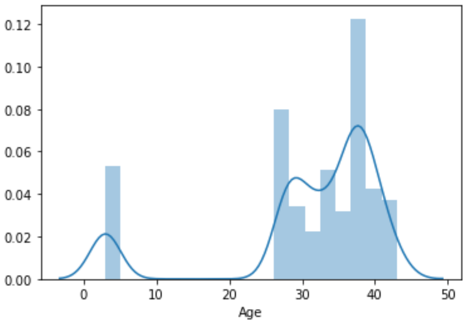
Percentage distribution-based Age group

From the age 30 to 40 maximum percentage of absenteeism. 30 to 35 bracket is almost equal to 35-40 bracket. Hence there is an equal distribution in these age groups.

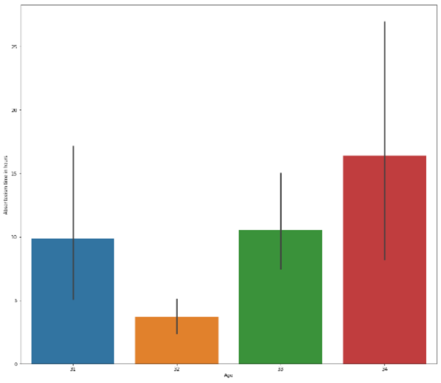




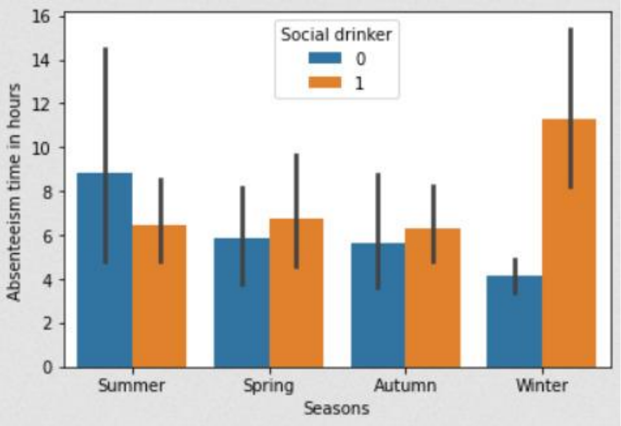
Also, from the age wise distribution it can be noticed that the data also contains mostly this distribution. There are few age group between 0 – 10 which should be considered as outliers.

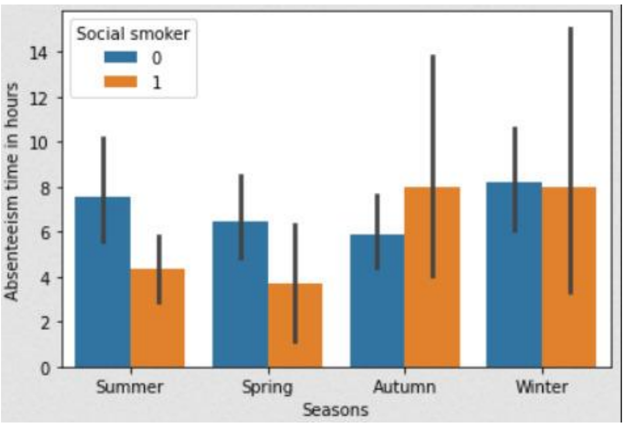


Further analysis showed age 34 had the highest among the 30-35 age bracket.



Patterns of social drinking/smoking people based on Seasons



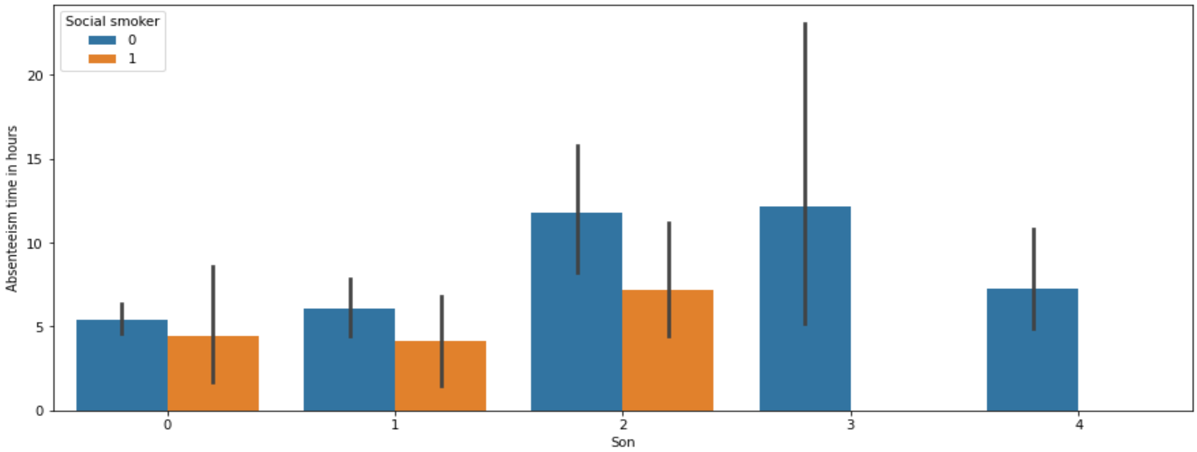


The data suggests that social drinkers are prone to take more leaves during winter when compared to all other seasons. Social Smokers tend to take more leaves during autumn and winter.

Patterns of social drinking/smoking people based on #Children

(a) Social Smoking:

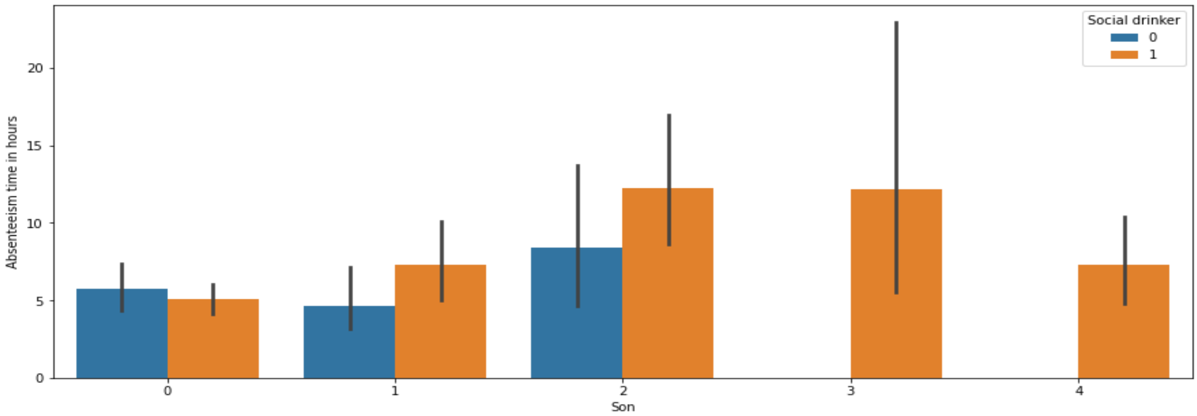
Non-Social Smokers are seen to take more hours of absenteeism from the graph shown below.



(b) Social Drinking:

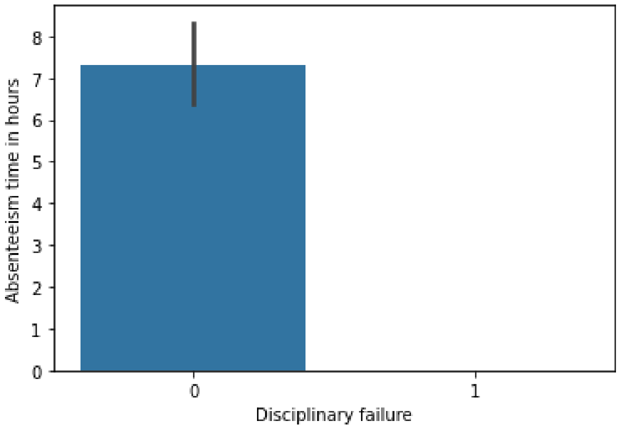
People who are involved in social drinking are generally seen to be absent for more number of hours.

It is noticed that a greater percentage of absenteeism hours is from people drinking socially(if they have more than 1 sons). Also, people with 3 or more sons are seen to drink socially.



Distribution based on disciplinary failure

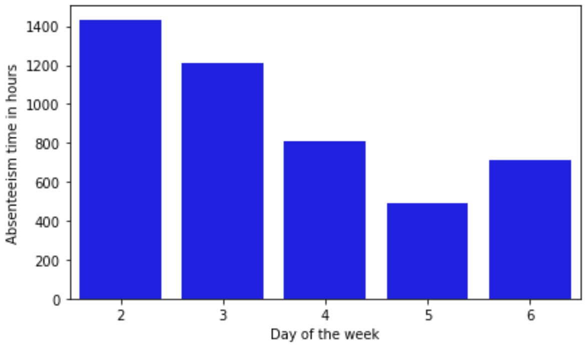
All the absenteeism was noticed with disciplinary failure 0.

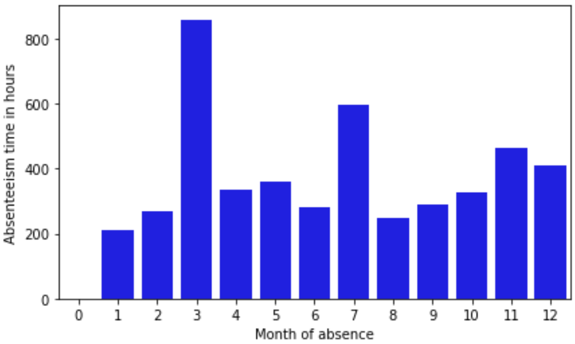


Time based on the day/month of absence

People take most time off at the beginning of the week. They gradually become more consistent as the week progresses, but as the weekend nears, people again take lots of absentee time.

People take vacations in the month of march(summer) and july, as there is a sudden rise in the absenteeism time as shown in the bar chart below.

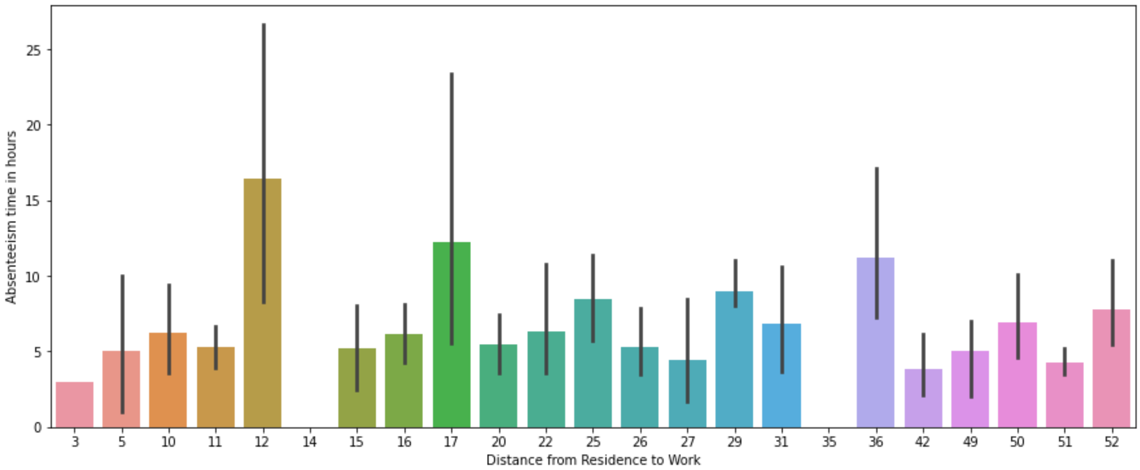




Time based on the distance of residence to work

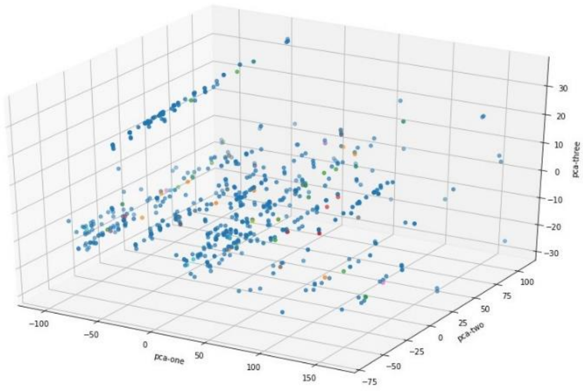
From the graph, it is clearly shown that hours of absenteeism are variating uniformly over the entire range of distances.

Although, the variance is more in the hours of absenteeism for people who live close to the place of work.

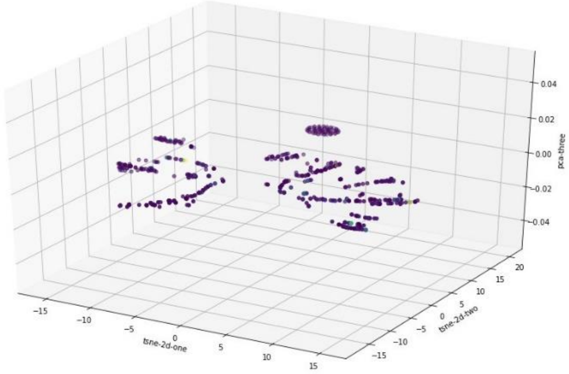


Dimensionality Reduction Results

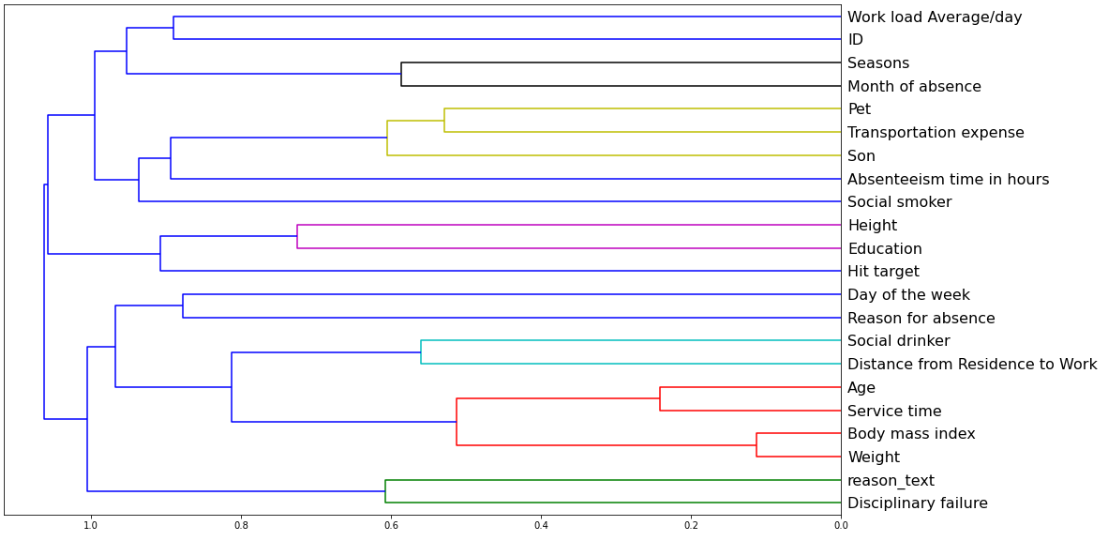
PCA

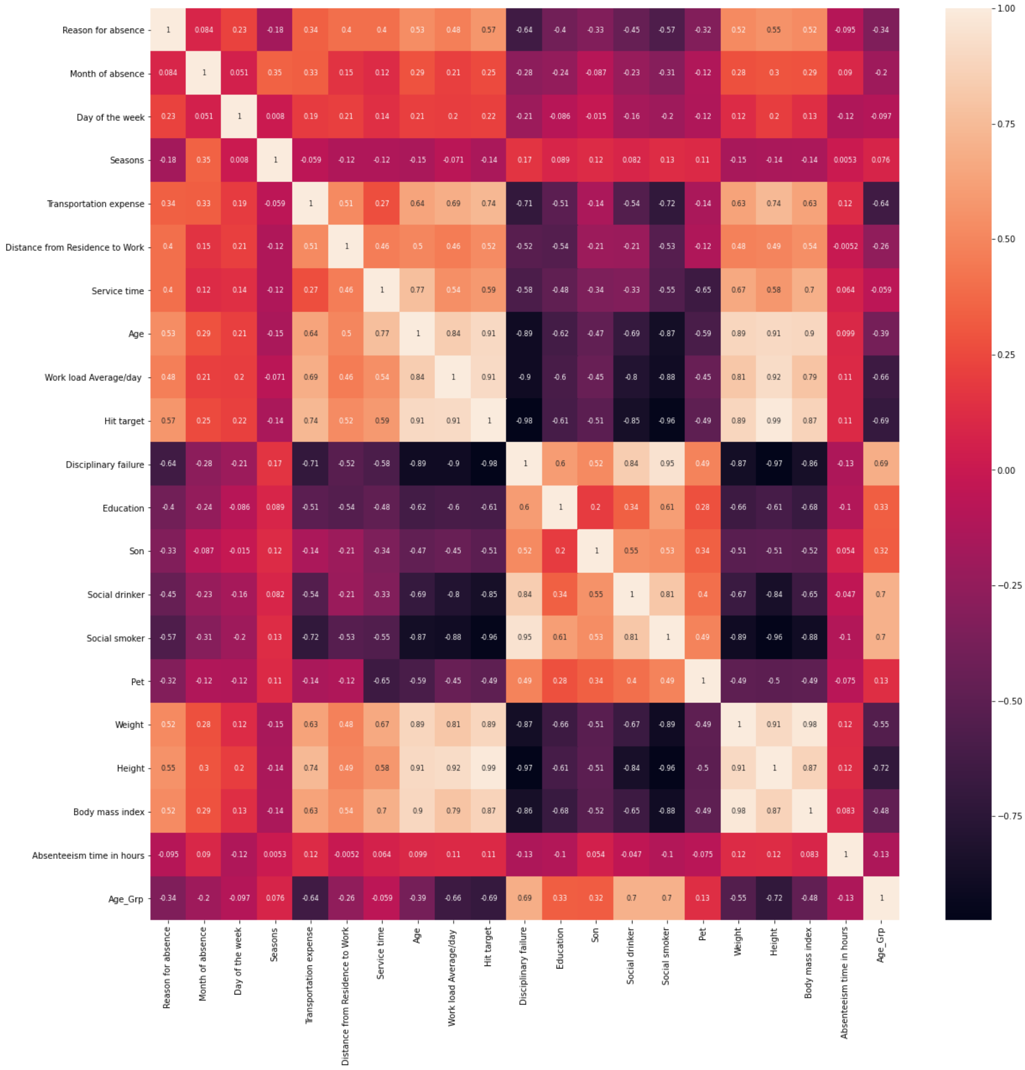


tSNE



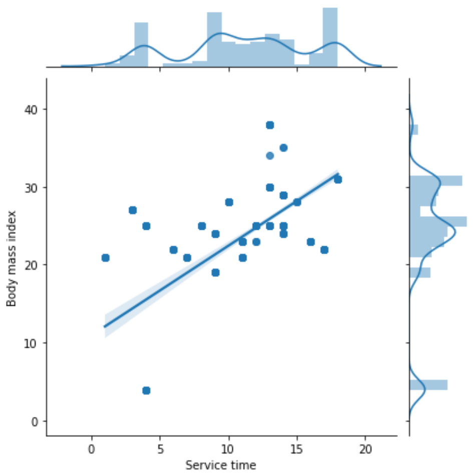
Correlation plots

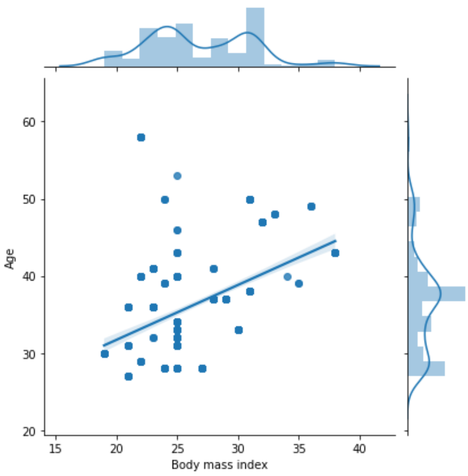




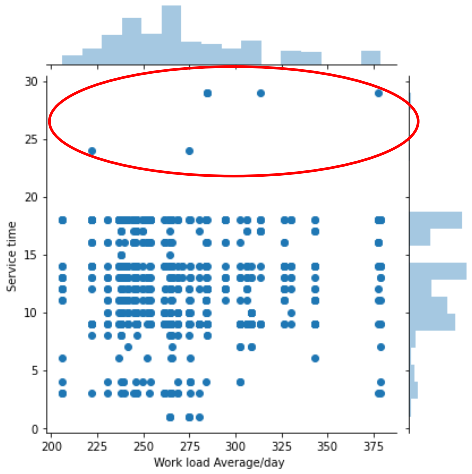
Other noticeable facts

As there is increase in the service time the BMI of the person generally increases and since BMI is closely co-related to weight it. Also, a general trend is that BMI increases with Age.

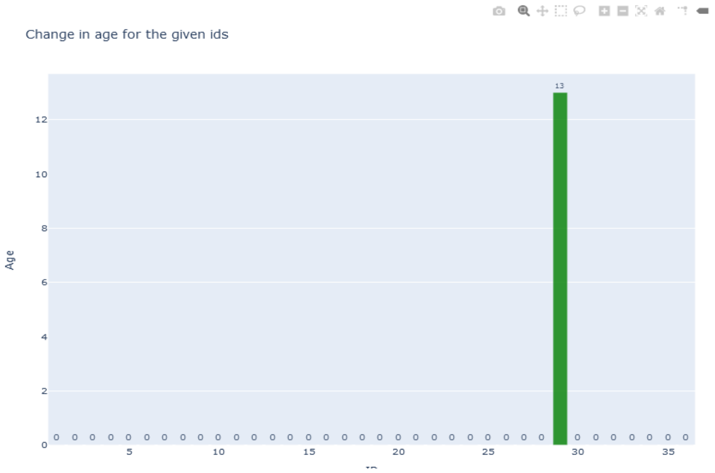




Also, we can say that the workload is evenly distributed and there is no preference based on service time. Although there can be noted a few outliers at the top part of the graph.

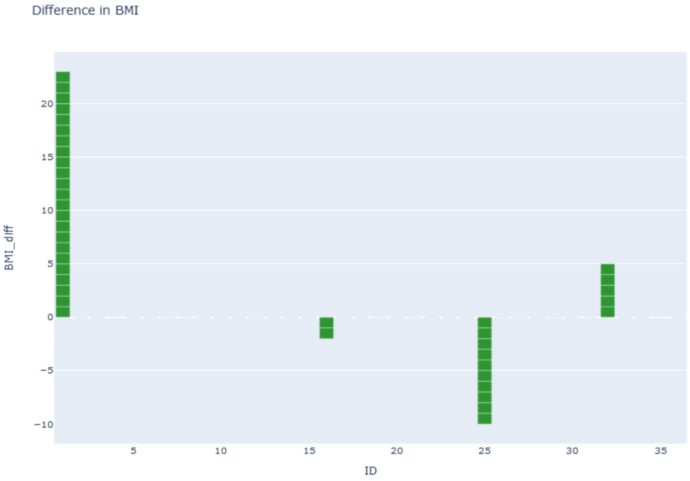


Anomalies detected in the Data



1. The data provided was for three years and in three years the age of the employees should have changed by 3 only but this was not the case observed in the dataset. For the employee id 29 there was a 13 year difference noticed from the data and also all other employee age never changed.

2. Also, for few of the employees there was a difference in the BMI calculation and also the difference was greater than 1(difference of 1 could represent the rounding off errors) which could be an error in the data.



Footnote: Plots were made using Python using libraries Seaborn, plotly-express, matplotlib and Scipy.