Exploratory Data Analysis

Cascade Cup - Round 3

Team: **ML Hacksters**

Members: Nippun Sharma (Sophomore - B.Tech Computer Science, IIT Mandi)
Ph. No. : 7701941189, Email : inbox.nippun@gmail.com
Naveen Saisreenivas (Sophomore - B.Tech Computer Science, IIT Mandi)
Ph. No. : 6230326108, Email : naveensaisreenivas@gmail.com

Introduction and Approach

This dataset contains absenteeism records for a courier company. It is a little different from the usual datasets because the **ID**'s are repeating (which signifies that an employee is absent more than once).

The given dataset is a record-type dataset, i.e. it consists of records of absenteeism. There are majorly 2 groups of features - **features dependent on ID** and those **independent of ID**. For eg, different records with the same **ID** will have the same values for features like ***Age, Height, Weight, BMI,*** etc whereas they will have different values for features like ***Month of absence, Day of absence, Hit target,*** etc.

Due to the different nature of these features, it will be **incorrect** to directly apply **analysis** on all the **features together.** Therefore, we have divided the data into two different sub-datasets - one containing the information unique to each employee which we will refer to as **Information Dataset** (dependent features like ***Age, BMI***) and the second which contains the records for absenteeism which we will refer to as **Record Dataset** (independent features like ***Month of absence, Reason for absence***). We have first done analysis separately on each sub-dataset and then we have tried to find some relationship between the two.

Data Cleaning

We observed that 44 records have ***Absenteeism time in hours*** as 0, out of which 40 records had ***Reason for absence*** as 0 and 3 records have ***Month of absence*** as 0. Hence, we decided to drop these records. The last section of the report contains analysis on the 40 dropped records that had ***Reason for absence*** as 0 and also contains some interesting facts about why we dropped them!

Feature Analysis

The following are **numerical features** -

- ➢ Transportation expense
- ➢ Distance from Residence to Work
- ➢ Service time
- ➢ Age
- ➢ Work load Average/day
- ➢ Hit target
- ➢ Son
- ➢ Pet
- ➢ Weight
- ➢ Height
- ➢ Body mass index
- ➢ Absenteeism time in hours

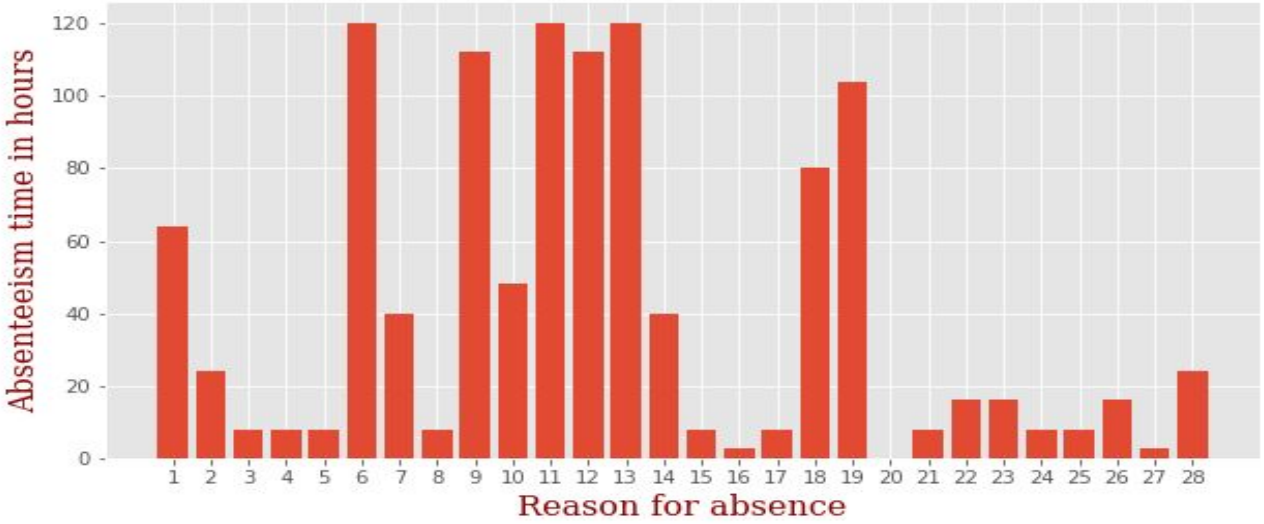
The following are **categorical features** -

- ➢ Reason for absence
- ➢ Month of absence
- ➢ Day of the week
- ➢ Seasons
- ➢ Disciplinary failure
- ➢ Education
- ➢ Son
- ➢ Social drinker
- ➢ Social smoker

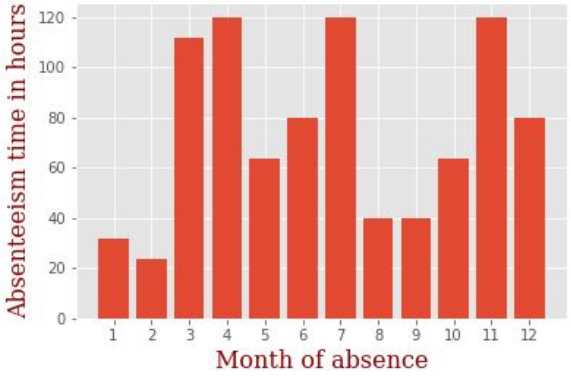
The following are features that are **independent of ID** -

- ➢ Reason for absence
- ➢ Month of absence
- ➢ Day of the week
- ➢ Seasons
- ➢ Work load Average/day
- ➢ Hit target
- ➢ Absenteeism time in hours
- ➢ Disciplinary failure

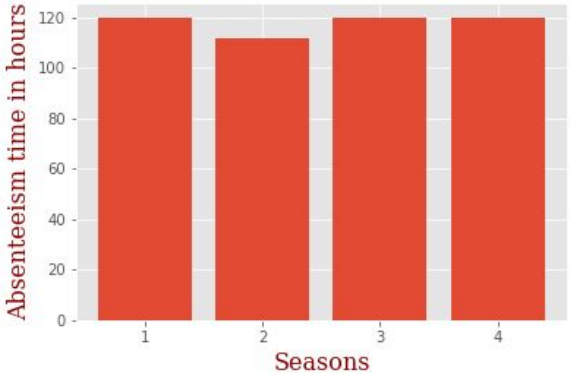
Bar Graphs



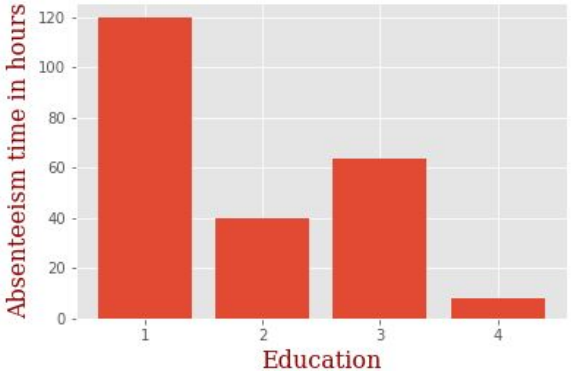
Our first analysis will be of the feature **Reason for absence**. Most employees are absent because of diseases of the nervous system (6), diseases of the digestive system (11) (diarrhea, constipation etc.) and diseases of the musculoskeletal system (13) (sprain, fractures etc.).



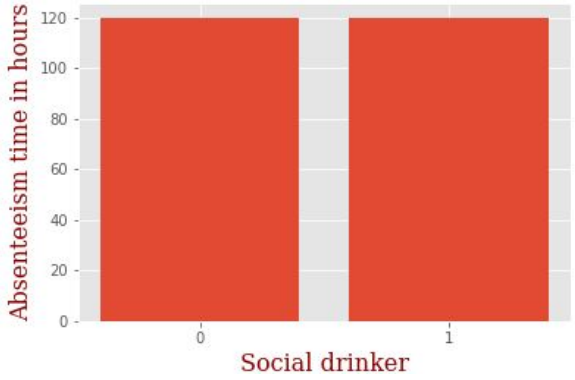
We can see that employees are most often absent during the Summer (March, April), July and November



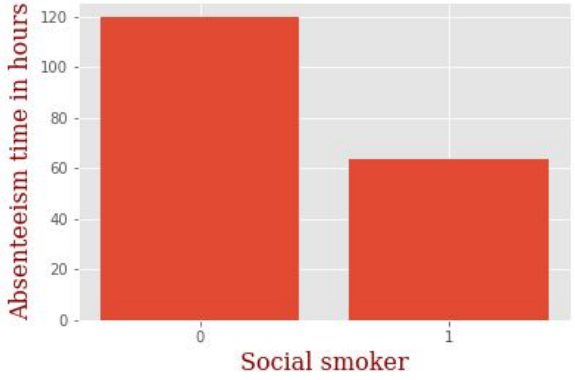
Employees are absent for almost equal no. of hours in all the seasons, except Autumn where it is slightly less.



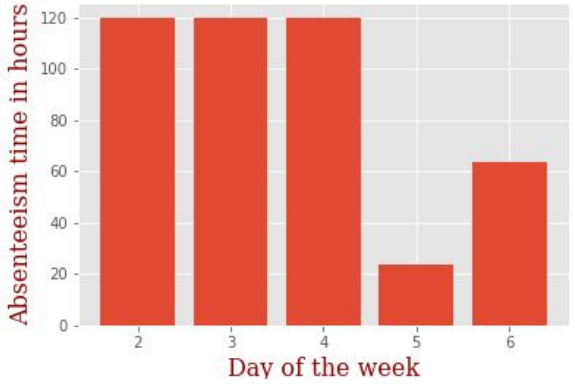
It is quite evident from the graph that employees who don't have a good education are absent more often. Either the company must counsel the employees who are high-school graduates, or they must stop hiring them to increase productivity in work.



Whether an employee drinks or not, it does not affect their work productivity in this case



Contrary to what we'd expect, an employee is more likely to be absent if he/she does not smoke than if he/she smokes!



Another interesting thing to note is that employees are absent mostly on Mondays, Tuesdays and Wednesdays!

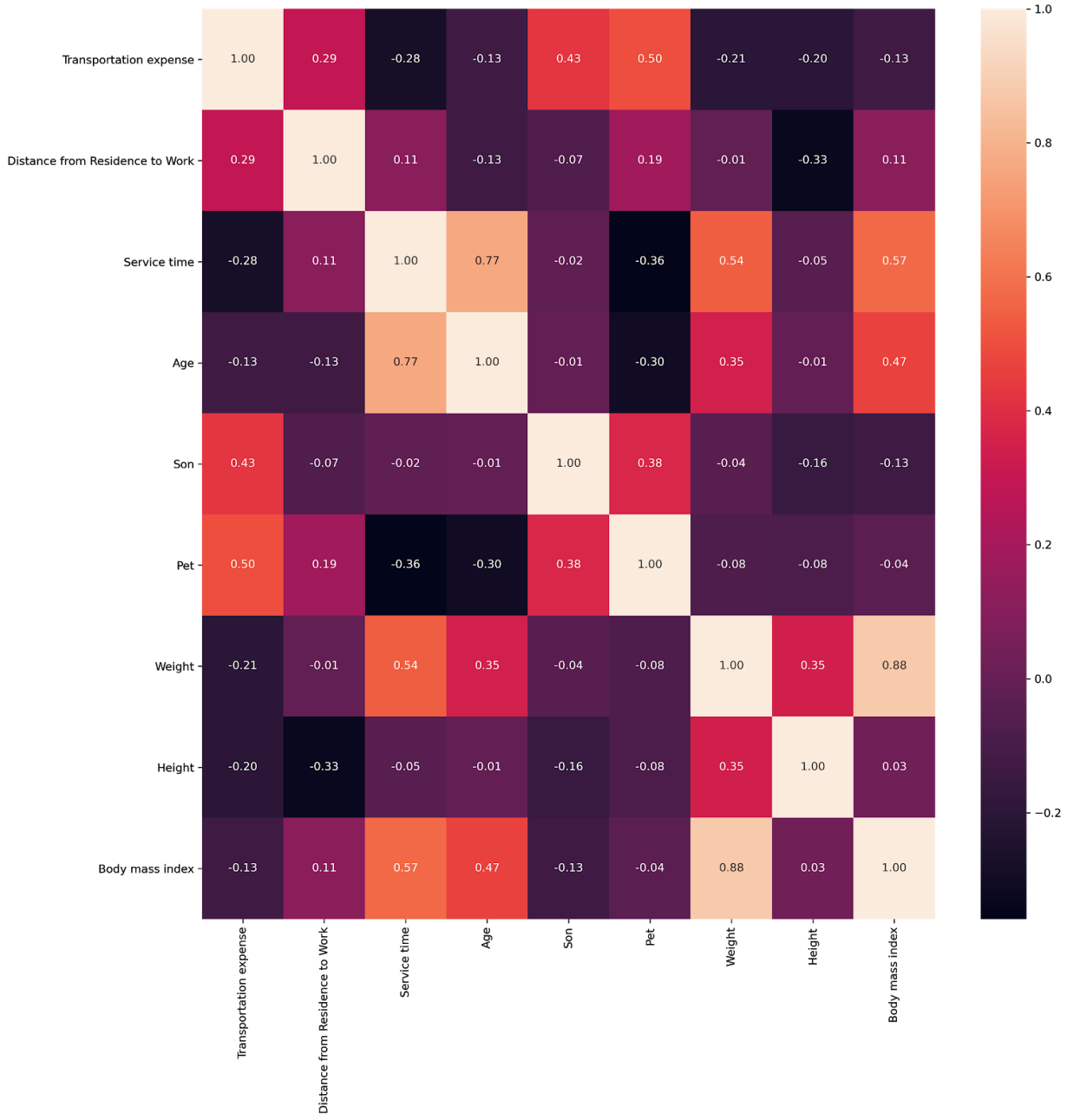
Correlations

1. Spearman's Correlation Coefficient (for numerical features **dependent on ID**)

   We went with the Spearman correlation heatmap because it can capture nonlinear monotonic relations between features whereas the Pearson correlation can capture linear relations only. Many features have ordinal values such as ***Age, Son, Pet, Service time*** and using Spearman correlation might work better than Pearson correlation.

   Some observations from the correlation heatmap:

   1. ***Weight*** is highly positively correlated to the ***Body mass index*** (which should be expected because BMI is directly proportional to the weight).
   2. ***Age*** and ***Body mass index*** have a slight positive correlation (which is because people tend to gain more mass or get "fatter" as their age increases, excluding very old people).
   3. ***Service time*** is highly positively correlated to the ***Age***. This is because older people have already spent more time working with the company as compared to the younger ones.
   4. ***Distance from residence to work*** and ***Transportation expense*** also have a low positive correlation. This might be because people coming from farther places have to pay more bus fares while getting to the workplace.

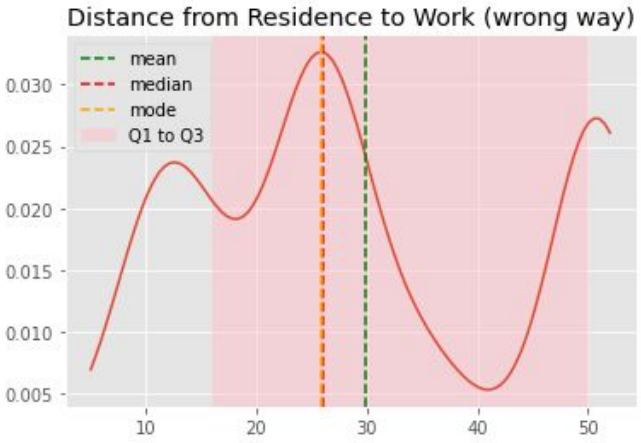


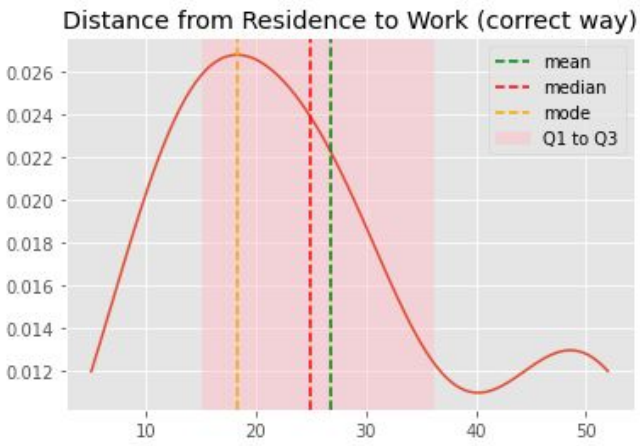
2. Spearman's Correlation Coefficient (for numerical features **independent of ID**)

Density Plots

We tried to estimate the densities of each of the **dependent numerical features** using **Gaussian Kernel Density Estimation** and found that some numerical features follow a distribution close to multimodal gaussian distribution within a certain range. Let us first take a look at the KDE plot for **Distance from residence to work**.

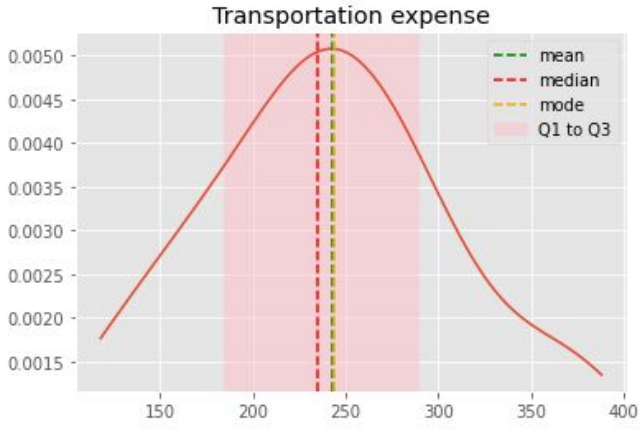
To explain why our approach is better than the usual approach, we have included two KDE plots. The left plot is the KDE plot for **Distance from residence to work** on the entire (original) dataset (where we have repeating values due to the same **ID**), whereas the one on the right is the KDE plot for **Distance from residence to work** from the **Information Dataset** (as defined on Pg-1). We expect the plot to be close to a gaussian distribution (**Central Limit Theorem**), but the plot on the left is clearly not gaussian, thus proving the fact that analysis cannot be directly applied on the original dataset.

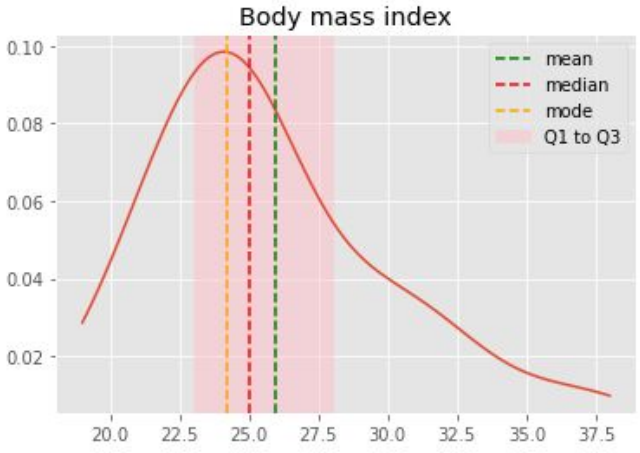


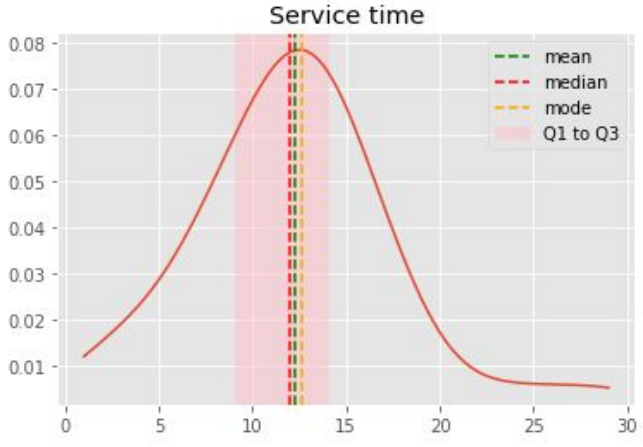


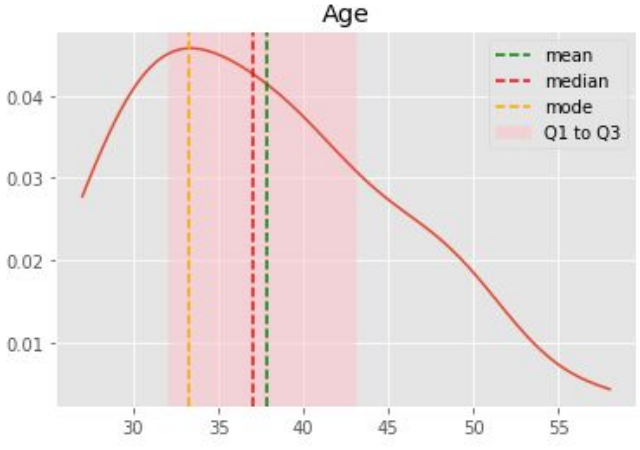
(Note: An important point to remember is that many features in this dataset have repeated values due to multiple records of absence of the same employee and while calculating the KDE we have to consider the values corresponding to each employee once only.)

As we can see from the above plot that the feature is close to a gaussian distribution. Let us take a look at some more features now:









We can clearly see that many numerical features follow a distribution close to Gaussian distribution.

t - tests

A. t-test between ***Absenteeism time in hours*** and ***Distance from Residence to Work***
Splitting the samples into two groups, less than or equal to 26kms to work and greater than 26km to work. Let's assume a significance value of α = 0.05.

Null Hypothesis : H_0 - $\mu_1 \geq \mu_2$
Alternate Hypothesis : H_a - $\mu_1 < \mu_2$

where μ_1 - Average absenteeism time for employees who travel ≤ 26 kms
μ_2 - Average absenteeism time for employees who travel > 26 kms

p-value for one-sided test p **= 0.08794375522124419**
As $p > \alpha$, we **fail to reject** the Null Hypothesis.
∴We can say that employees with lesser commute to work have on average higher absenteeism time.

B. t-test between ***Absenteeism time in hours*** and ***Transportation Expense***
Splitting the samples into two groups, less than or equal to 225 units for transport and greater than 225 units. Let's assume a significance value of α = 0.05.

Null Hypothesis : H_0 - $\mu_1 \leq \mu_2$
Alternate Hypothesis : H_a - $\mu_1 > \mu_2$

where μ_1 - Average absenteeism time for employees who pay ≤ 225 units
μ_2 - Average absenteeism time for employees who pay > 225 units

p-value for one-sided test p **= 0.018470626902858084**
As $p < \alpha$, we **reject** the Null Hypothesis.
∴We can say that employees with lesser cost of transportation on average have higher absenteeism time.

C. t-test between ***Absenteeism time in hours*** and ***Work load Average/day***
Splitting the samples into two groups, less than or equal to 264.249 units of work on average and greater than 264.249 units. Let's assume a significance value of α = 0.05.

Null Hypothesis : H_0 - $\mu_1 \leq \mu_2$
Alternate Hypothesis : H_a - $\mu_1 > \mu_2$

where μ_1 - Average absenteeism time for employees who work ≤ 264.249 units on average
μ_2 - Average absenteeism time for employees who work > 264.249 units on average

p-value for one-sided test p **= 0.09568737102988363**
As $p > \alpha$, we **fail to reject** the Null Hypothesis.
∴We can say that employees with higher workload have on average higher absenteeism time.

χ^2 tests

We have converted ***Absenteeism time in hours*** to a categorical variable by binning it as follows -

$$0 \leq x \leq 2$$
$$2 < x \leq 8$$
$$8 < x \leq 24$$
$$24 < x \leq 48$$
$$x > 48$$

where x is the Absenteeism time in each record.
Then, for each pair of categorical features, we performed the test using the following hypotheses -

Null Hypothesis : H_0 - $Feature_x$ and $Feature_y$ are independent
H_a - $Feature_x$ and $Feature_y$ are dependent

For all the tests, we have assumed a significance level of α = 0.05.
Note - p-values along diagonals are not considered for analysis as this would mean that we are testing a feature with itself.

A. χ^2 test among categorical variables independent of ID
    1. χ^2 Statistic Value Table

Features	Reason for absence	Month of absence	Day of the week	Seasons	Absence binned
Reason for absence	18096	505.463	128.532	243.511	368.044
Month of absence	505.463	7656	39.215	1725.149	54.562
Day of the week	128.532	39.215	2784	12.47	44.22
Seasons	243.511	1725.149	12.47	2088	17.148
Absence binned	368.044	54.562	44.22	17.148	2784.00

2. p-values

Features	Reason for absence	Month of absence	Day of the week	Seasons	Absence binned
Reason for absence	0.000	0.000	0.052	0.000	0.000
Month of absence	0.000	0.000	0.677	0.000	0.132
Day of the week	0.052	0.677	0.000	0.409	0.000
Seasons	0.000	0.000	0.409	0.000	0.144
Absence binned	0.000	0.132	0.000	0.144	0.000

We can say that all pairs of features that have a p-value lesser than 0.05 are dependent and all pairs of features that have a p-value greater than 0.05 are independent. Hence, these 5 pairs of features are dependent -

- Month of absence and Reason for absence
- Seasons and Reason for absence
- Seasons and Month of absence
- Absence binned and Reason for absence
- Absence binned and Day of the week

B. χ^2 test among categorical variables dependent on ID

1. χ^2 Statistic Value Table

Features	Education	Social drinker	Social smoker
Education	24	2	1.5
Social drinker	2	1	0
Social smoker	1.5	0	1

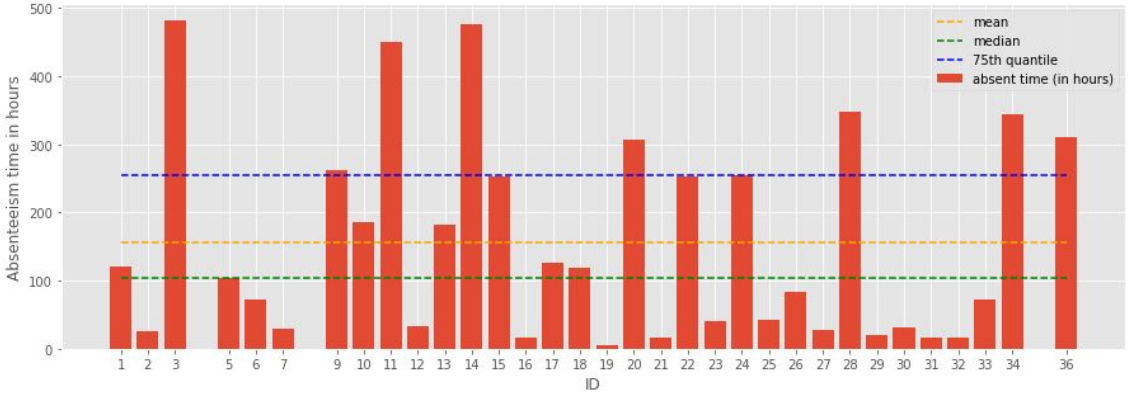
2. p-values

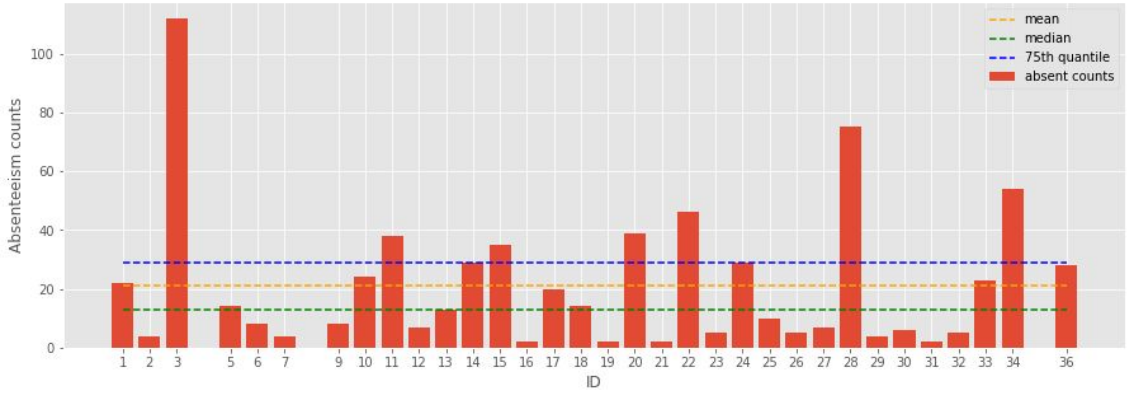
Features	Education	Social drinker	Social smoker
Education	0.004	0.572	0.682
Social drinker	0.572	0.317	1.000
Social smoker	0.682	1.000	0.317

As all the p-values are above 0.05, we can conclude that all of the above features are independent of each other. This means that having a better education doesn't mean that the employee will not smoke / drink. This also means that an employee who smokes does not necessarily drink and vice-versa.

ID Based Analysis

The dataset contains absence records of 36 different employees. In this section we will try to find patterns and make inferences from the records for each employee.
Note: From here on, each person will be referred to as **Pi (Person i)** where **i** is the **ID** of that person.
We will first plot the **Absenteeism time in hours** (total time of absence) and **Absenteeism count** (total count of absence) for each person :





(Note: **ID**'s **4,8,35** are not shown in the graphs as they are removed. All of them have **Absenteeism time in hours** as 0)

- In each of the bar plots we have plotted the **mean** and **median** also. To analyse which employees are absent more often and for a longer duration, we will use the **75th quartile as the threshold**.
- From the plots we see that **P3, P9, P11, P14, P20, P28, P34, P36** have **more absenteeism time** than others.
- Also, **P3, P11, P15, P20, P22, P28, P34** have **more absence counts** than others.

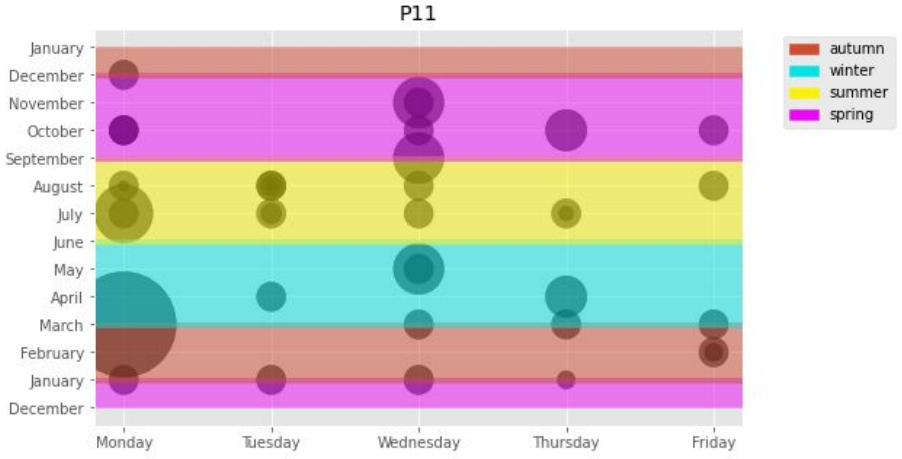
From the above observations we see that **P3, P11, P20, P28, P34** are **common defaulters**. They are absent more often and for a longer time.

We will now try to analyse the ***"time"*** pattern of the absences i.e. the **days** or **months** when an employee is absent more often. We will analyse this pattern using a **bubble chart** where the **size** of each "bubble" represents the **time of absence** and the **concentration** of "bubbles" represent the **frequency / counts of absence**.

The **horizontal axis** represents each **day** of the week and the **vertical axis** represents the **months**.
We have also colored the background using a **gradient** in which each **solid color** represents a **season**.

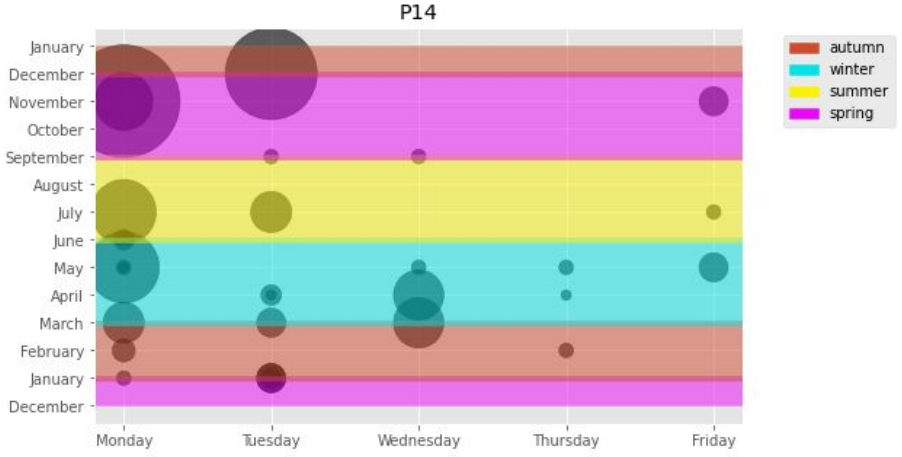
(We determined which months are related to which seasons from the dataset itself. For eg:- for season 1 or summer season we found all the unique values of the **Month of absence** feature where the season is equal to 1 and it gave us the values 6, 7, 8, 9 i.e. June, July, August, September correspond to the summer season. Also note that some months belong to two seasons and they are taken as the ***transition*** from one season to another and this is clearly visible in the plots below.)

Let us take a look at **P11**'s absence pattern:



From the above plot we can see that **P11** is absent most frequently on Wednesday. Also, most of the time, he/she is absent during the Summer season.

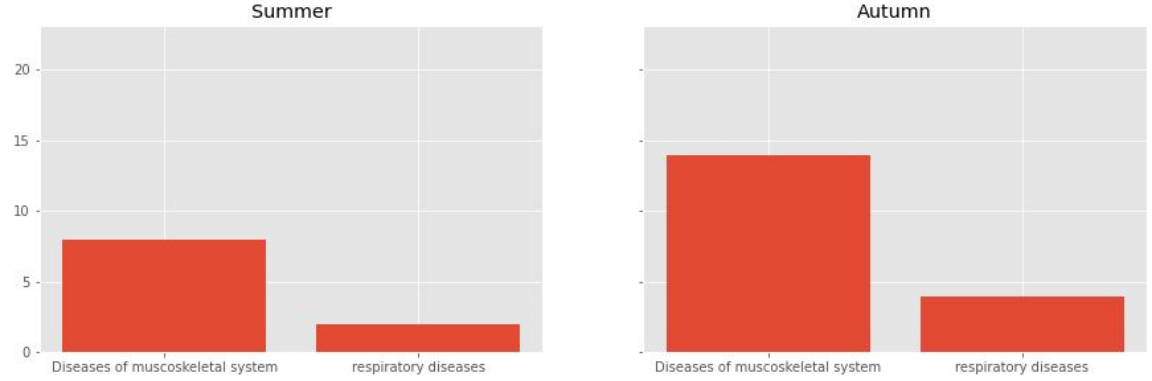
Just for fun, let us also analyse **P14**'s absence pattern:

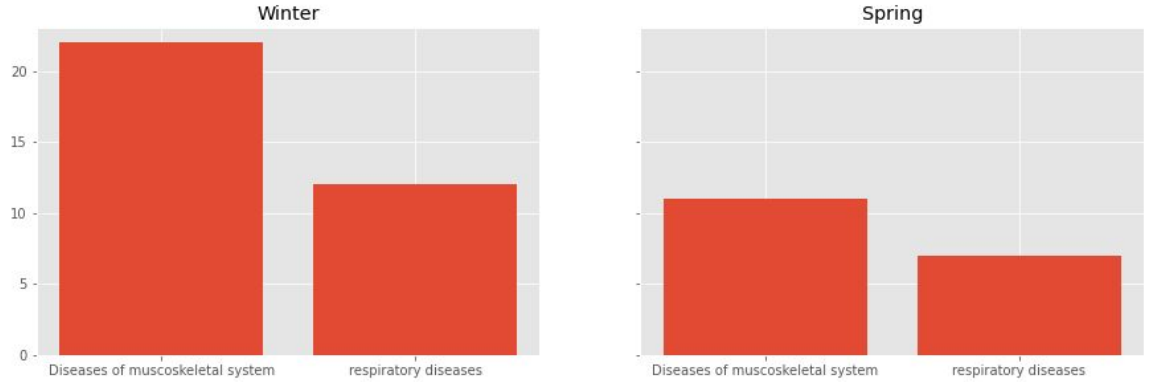


From the above plot we can see that **P14** is absent most frequently on Monday. Also, most of the time, he/she is absent during the Winter season.

Interesting Facts & Findings

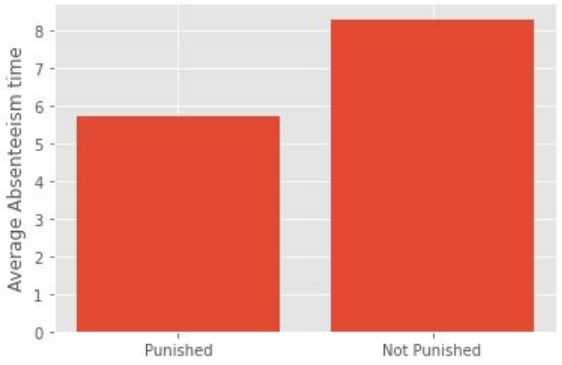
1) We observed that the musculoskeletal diseases (sprain, fractures, etc.) are more common in the Winter season (almost 100% more from Summer and 25% more from Autumn) as compared to any other season. The same is the situation with respiratory diseases (cold, flu etc.). They are more common in the Winters.





These observations suggest that the employers should take better care of the employees in the Winter season as they are more prone to common health related issues during this period. A decreased amount of workload might help the employers in reducing the chance of these diseases and hence reduce absenteeism.

2) We will now analyse the **Disciplinary failure** feature.
All entries that have **Disciplinary failure** value as 1 also have their **Absenteeism time in hours** as 0. We think that this might be an entry added to the record as a **punishment** for indiscipline shown by the employee at the workplace.
We will now analyse the absenteeism time of employees who have been punished and those who have not been punished:



In the above figure we plot the average absenteeism time of employees who are punished (or have **Disciplinary failure** as 1) at least once and employees who are not punished (or have **Disciplinary failure** as 0). We can clearly observe that employees who have been punished at least once have a lower **Average Absenteeism time** as compared to non-punished employees. We can say that taking some action against indiscipline at the workplace makes the employees follow rules even with greater caution!

3) Let us take a look at the **Transportation expense** vs **Absenteeism counts** for the employees. We found that the **Transportation expense** and **Absenteeism Counts** have a Spearman correlation of ***-0.27522,*** which signifies that many employees who have a lesser transportation expense are absent more often!

4) Let us also take a look at **Distance from Residence to Work** and **Average Absenteeism time** for the employees. We found that the Spearman correlation of these features is ***-0.3402***, which signifies that employees who stay closer to the workplace have more absenteeism time!

Points 3 and 4 imply that employees staying closer to work are absent more often. It might be the case that an employee comes late to work or goes home early because he/she stays closer.

Conclusion

Due to a limit on the number of pages, we couldn't include many other methods which would provide insights about the data. We could have **visualized** the data after **dimensionality reduction** using either **PCA** or **t-SNE**. Another possibility is that we could have grouped the employees using **clustering techniques** on the basis of the given features.

We thank **CnA, IIT Guwahati** and **Trell** for conducting such an amazing competition where results are not just based on models, but also takes into account data analysis, which is crucial for a data scientist. We have thoroughly enjoyed participating in the competition. Please hold more such competitions!

Note : For any clarification, please feel free to contact us via email or phone.