Applied Data Analytics (Feb 2019)

Msc. In Big Data Management And Analytics Assessment

Student Name : Siddhi Kate
Student Number : 2982006
Dataset : Absenteeism at work dataset

Data Analytics and Information Dashboards on the dataset in selected domain

Part I

Exploratory Analytics and Visualization of Data via Information Dashboard (40% of the module)

Work is an important aspect in one's life. On an individual level as well as on organization level, being able to efficiently follow the work schedule is utterly important. The organization's progress as well the individual's progress is at stake when it comes to working . All huge businesses are dependent on their employees for getting their work done and the employees are dependent on the organization for their livelihood. In order to ensure that this co-dependant relationship runs smoothly, it is important for employers to track the attendance of the workers, as well as explore all the factors that can affect this.

The dataset chosen has a set of attributes with **absenteeism at work** in hours being the target attribute. This database was been built by collecting data of employees from a courier company over the period from July 2007 to July 2010. Other than absenteeism , another important attribute is the reason for absence , which has a set of medical conditions that can cause absenteeism at work.

The dataset is currently available at UCI Machine Learning Repository which is a reliable source of datasets for Machine Learning and Intelligent Systems :

Link : https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

Source :

Creators original owner and donors: Andrea Martiniano , Ricardo Pinto Ferreira , and Renato Jose Sassi .

Universidade Nove de Julho - Postgraduate Program in Informatics and Knowledge Management.

Relevant Paper :

Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE.

Citation Request :

Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE.

Attribute Description :

1. **Individual Description (ID)** : It is an integer value and consists of unique value for every employee.
2. **Reason for absence :** It is an integer value , ranging from 1 to 28 .
   Values 1 to 21 stand for reasons that can be considered as International Code for Diseases(ICD). They are as follows :
   I Certain infectious and parasitic diseases
   II Neoplasms
   III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
   IV Endocrine, nutritional and metabolic diseases
   V Mental and behavioural disorders
   VI Diseases of the nervous system
   VII Diseases of the eye and adnexa
   VIII Diseases of the ear and mastoid process
   IX Diseases of the circulatory system
   X Diseases of the respiratory system
   XI Diseases of the digestive system
   XII Diseases of the skin and subcutaneous tissue
   XIII Diseases of the musculoskeletal system and connective tissue
   XIV Diseases of the genitourinary system
   XV Pregnancy, childbirth and the puerperium
   XVI Certain conditions originating in the perinatal period
   XVII Congenital malformations, deformations and chromosomal abnormalities
   XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
   XIX Injury, poisoning and certain other consequences of external causes
   XX External causes of morbidity and mortality
   XXI Factors influencing health status and contact with health services.

   Values 22 to 28 are as follows and cannot be considered as ICD :
   XXII Patient Follow-up
   XXIII Medical Consultation
   XXIV Blood Donation
   XXV Laboratory Examination
   XXVI Unjustified Absence
   XXVII Physiotherapy
   XXVIII Dental Consultation
3. **Month of Absence :** It is an integer value ranging from 1 to 12, referring to the month number (January – December) . It will help to understand what month has the greatest absenteeism.
4. **Days of the Week :** It is an integer value ranging from 2 to 6 corresponding as follows –
   Monday : 2
   Tuesday : 3
   Wednesday : 4

Thursday : 5
Friday : 6

5. **Seasons :** It is an integer value ranging from 1 to 4 , corresponding to following :
   Summer : 1
   Autumn : 2
   Winter : 3
   Spring : 4
6. **Transportation expense :** Integer Value , that will help us find out if such costs affect absenteeism at work .
7. **Distance from residence :** Integer values, approximate distance in kilometres.
8. **Service time** : Integer value, time in hours
9. **Age** : Integer value, specifying the age of the employee.
10. **Work load Average/Day** : Integer value
11. **Hit Target** : Integer value of percentage of the target achieved.
12. **Disciplinary failure** : yes = 1, no = 0
13. **Education** :Integer value,(high school (1), graduate (2), postgraduate (3), master and doctor (4))
14. **Son (number of children)** : Integer value
15. **Social drinker** : Integer value,(yes=1; no=0)
16. **Social smoker** : Integer value,(yes=1; no=0)
17. **Pet (number of pet)** : Integer value
18. **Weight** : Integer value
19. **Height** : Integer value
20. **Body mass index** : Integer value
21. **Absenteeism time in hours** : Integer value, (target variable)

While undertaking data analytics tasks, the following questions can be answered/ explored :

1.) Out of all possible variables, which is the variable that affects the most to the target variable?
2.) Does social life of an employee affects the absenteeism?
3.) What is the probability that employees, that live far from work place, tend to be more absent ?
4.) Is there any particular medical condition, that is seen widely as the reason for absence ?
5.) To what extent, does the physical fitness of an employee affect absenteeism ?
6.) Is there any relationship between seasons and reason for absence , which may lead to increase or decrease in absenteeism .
7.) How many times does a situation occurs where the reason for absence is unjustified ?
8.) Do employees with family responsibility have more absence hours ?
9.) How adversely does absence affect the hit target of the company ?
10.) What are the measures that organization should take in order to reduce absenteeism ?
11.) How can organization design their work schedule for eg. The number of hours an employee works depending on his physical fitness, family background and other factors ?
12.) What are the ways to find out actual reasons for unjustified absence ?

The above questions/issues are important for head department of the organization. Inefficient workforce can lead to loss of the organization. Hence it is important for organizations to analyse this kind of data, and search for causes of the absences and then design measures to minimize the situation as much as possible.

It can also help in finding out a major cause in a particular season or a month. Such as a communicable disease , that is causing most of the absences.

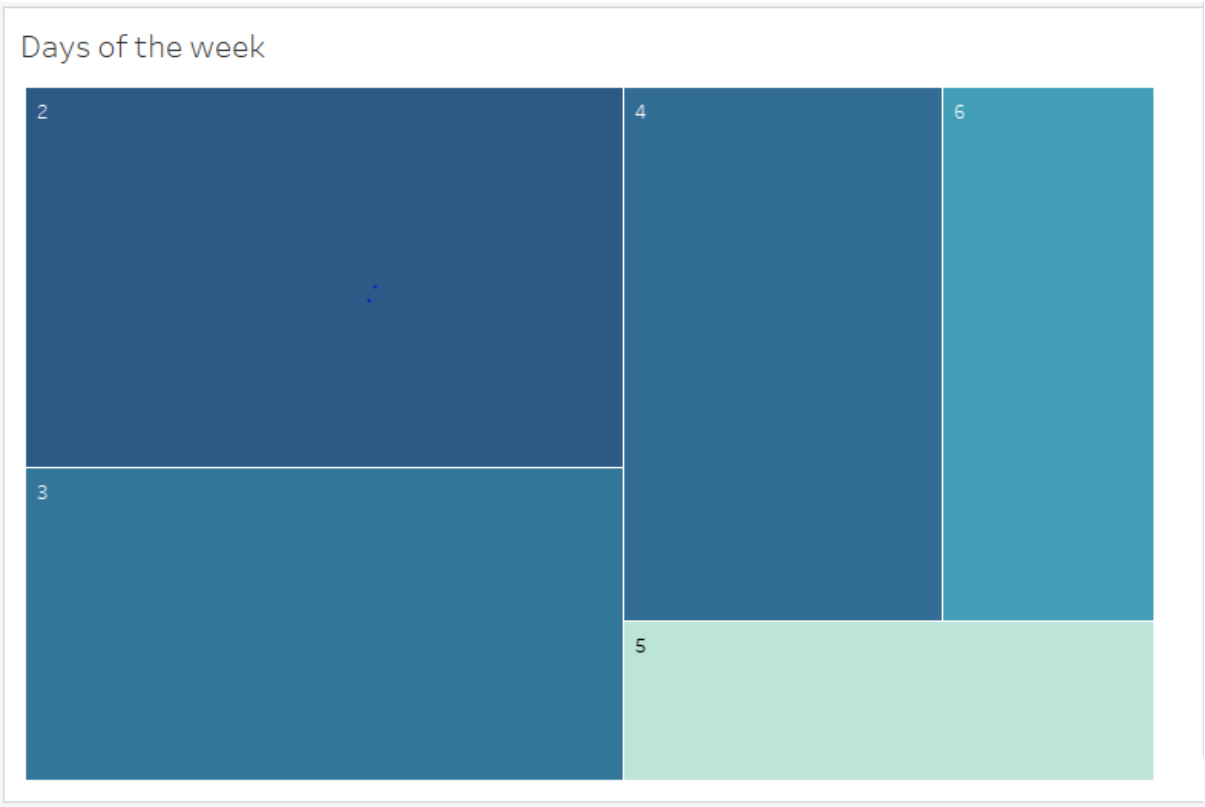
Depending on the above the problem statement can be defined as follows :

Analysing past records of employees based on various attributes such as social and family life, physical fitness, sickness or disease etc for predicting the causes of absences, followed by making decisions that can minimize the same depending upon the insights gained.

Exploratory Analysis using Information Dashboards :

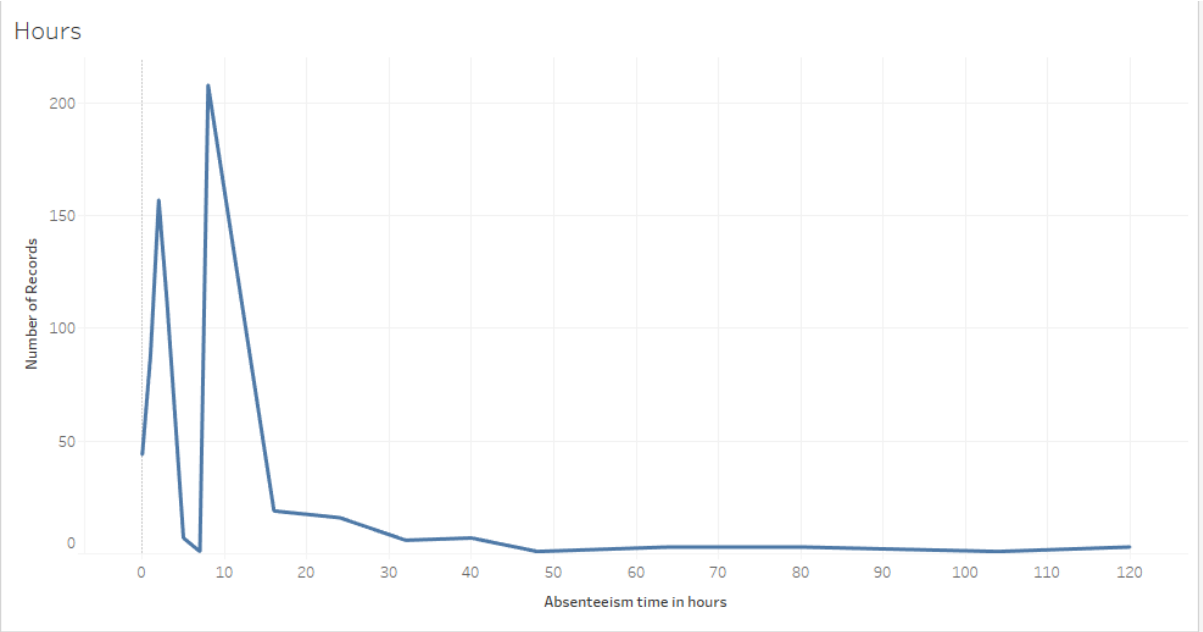
Dashboard 1 :
The first information dashboard is based on how the time aspect affects the absenteeism. Three separate graphs are created based on hours, days of the week and months of the year all mapped against individual records

Worksheet 1 :
This graph plots the number of records with absenteeism against the days of the week . Here Sum(absenteeism in hours is being used)



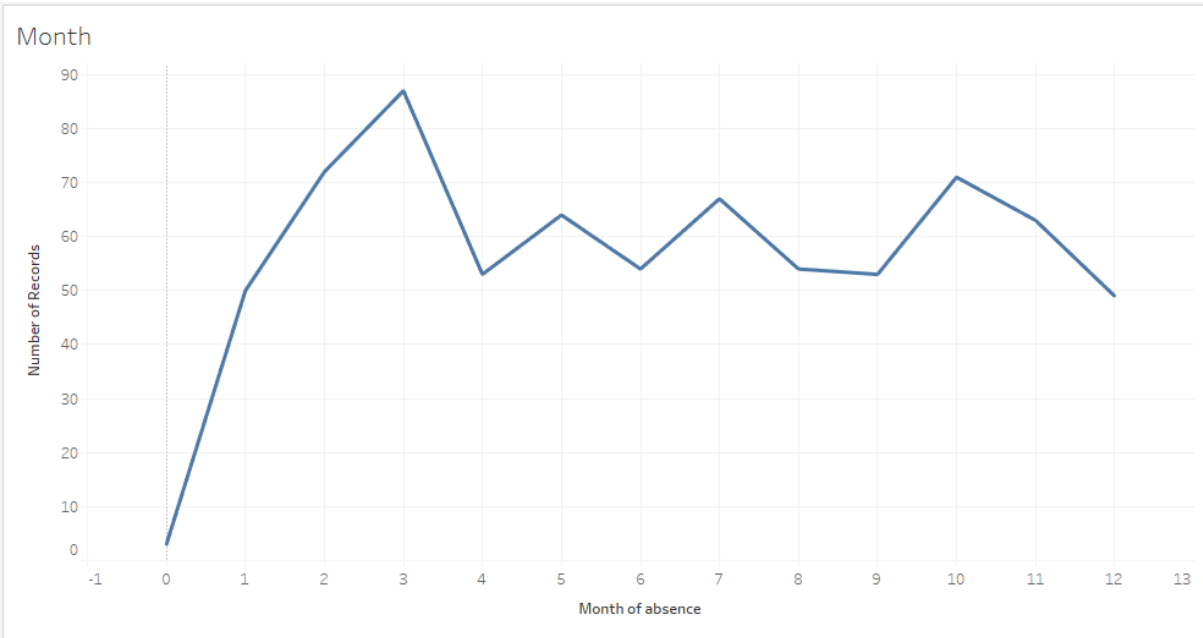
Worksheet 2 :

In this graph we plot the number of hours as a dimension i.e. to find the highest occurrence of average number of hours that an employee is absent from work .

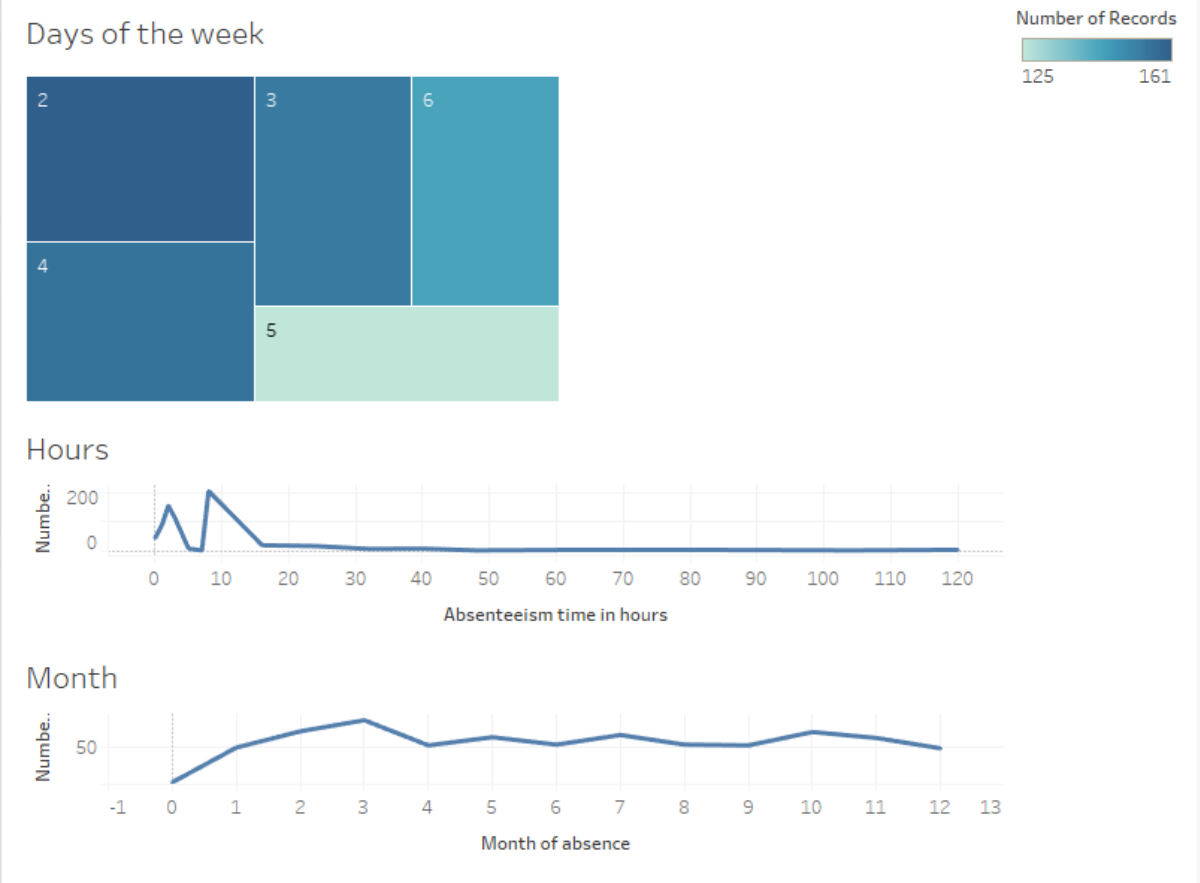


Worksheet 3 :

Here we plot Absenteeism based on the month of the year. The target attribute is selected as measure and mapped against number of records and month of absence.



Final Dashboard :
The final dashboard looks like the following image :

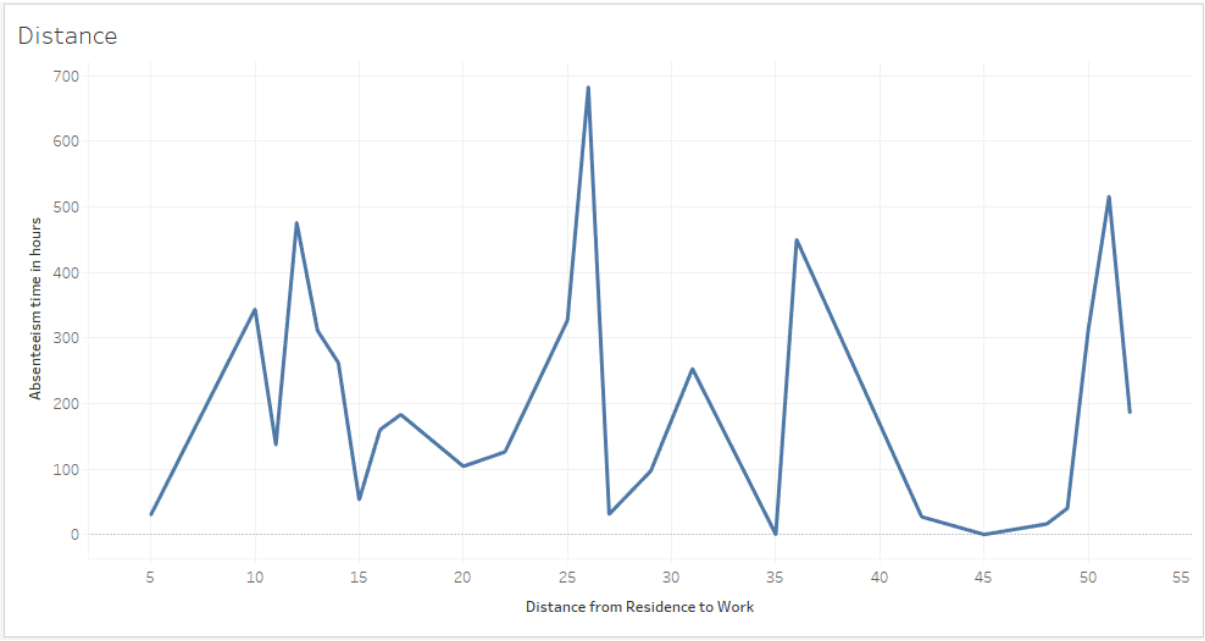


Insights gained from the above dashboard :

1) Monday, even though being the first day of the working week has maximum number of absences, whereas the least absences are seen on Friday which is the last day of the week.
2) The average number of hours that an employee tends is to be absent is approximately 10.
3) Maximum number of absences are seen in March and September.

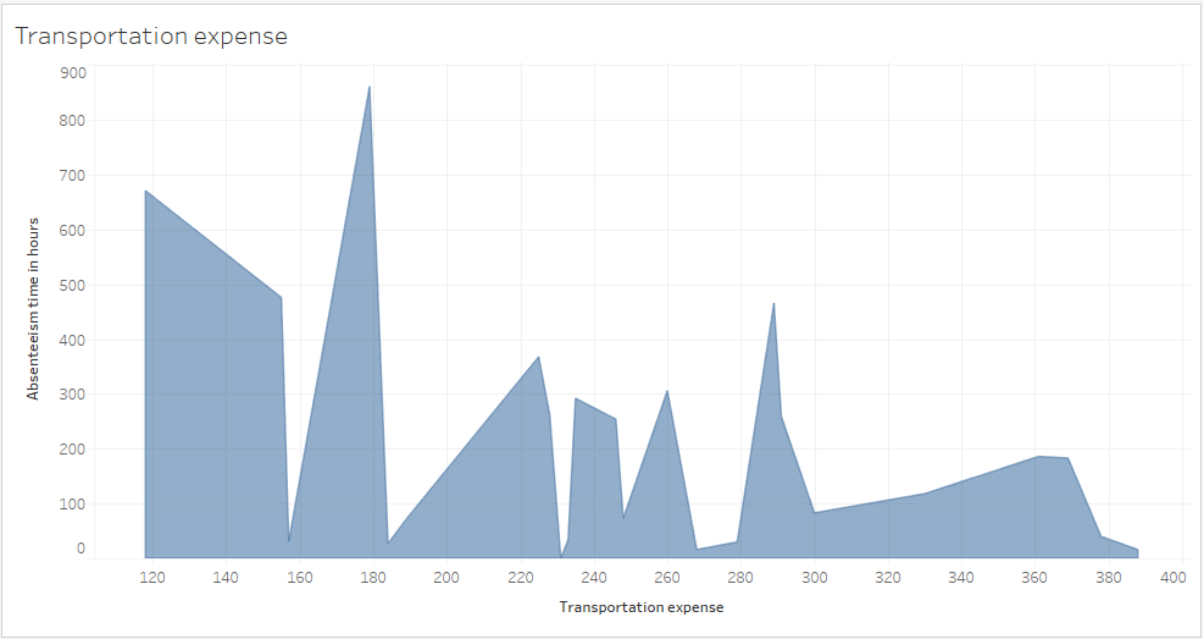
Dashboard 2 :
The second dashboard is based on how the travel aspect affects absenteeism . It also takes the workload average/day in order to see if increase/decrease in workload along with travel affects absenteeism.

Worksheet 1 :
A graph is plotted by using distance from residence in kms as dimension against the number of absenteeism in hours as measure . The following graph is obtained :



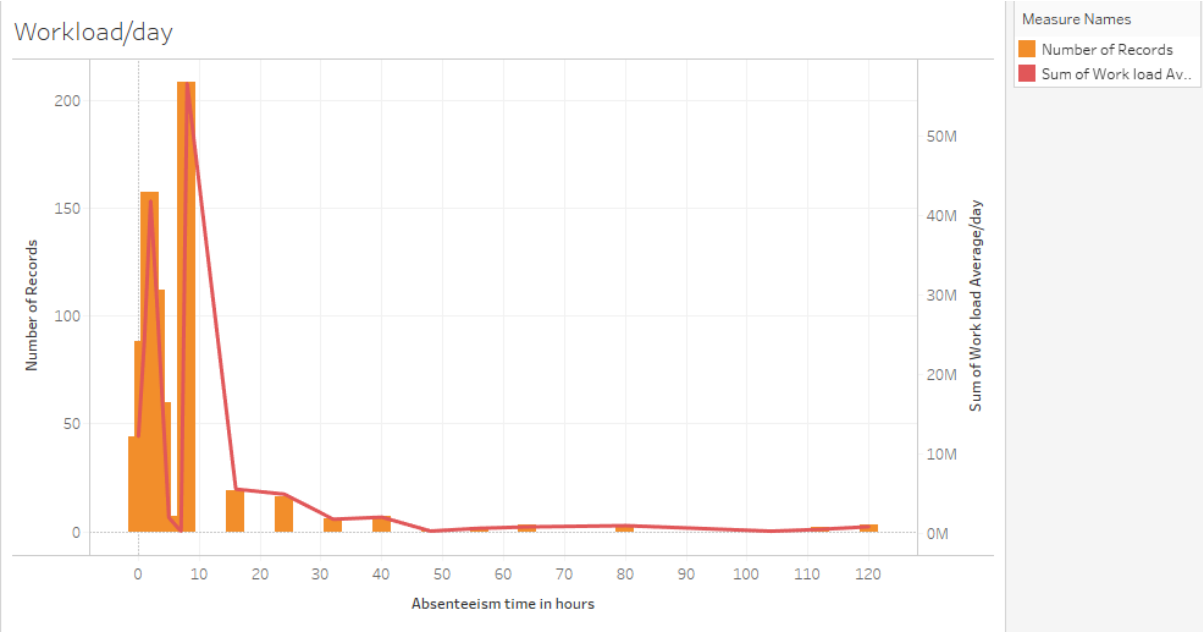
Worksheet 2 :

A graph is plotted by using transportation expense from residence as dimension against the number of absenteeism in hours as measure . The following graph is obtained :

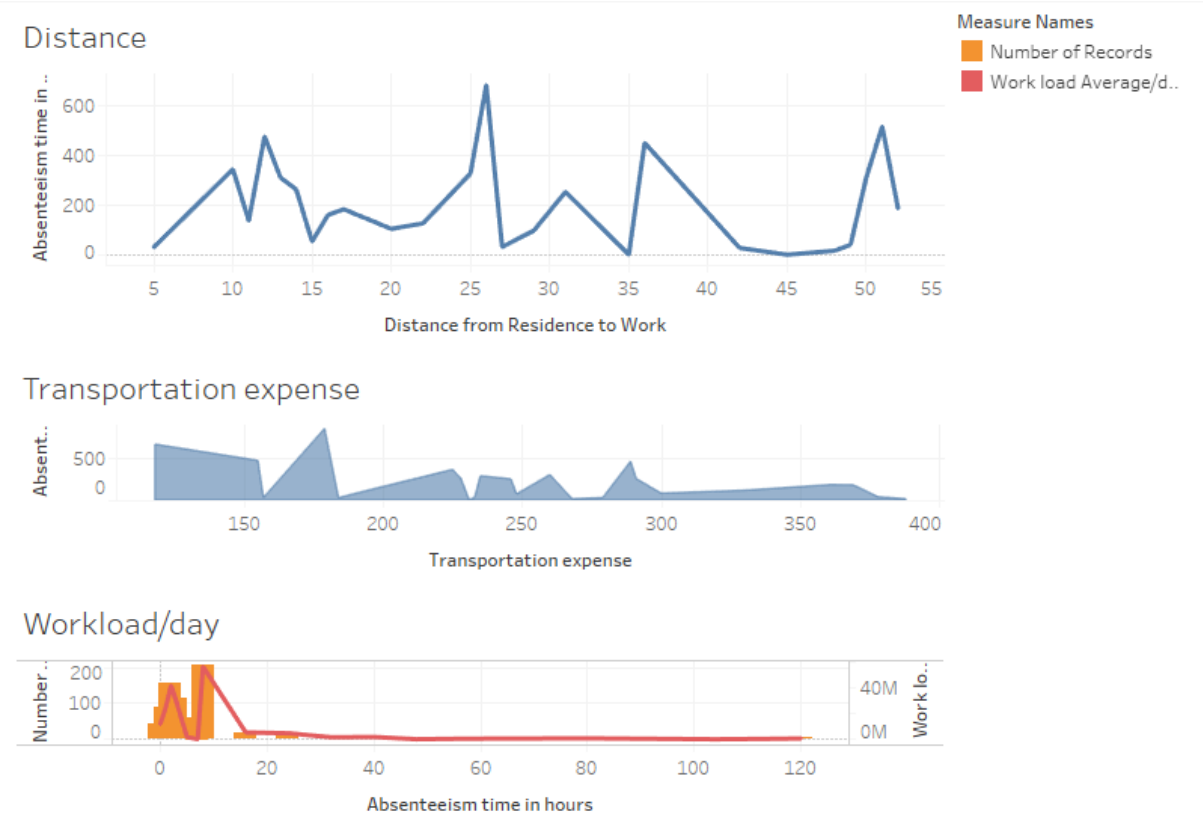


Worksheet 3 :

A graph is plotted by using Average workload/day as dimension against the number of absenteeism in hours as measure . The following graph is obtained :



The final dashboard :



Insights from the above dashboard :

1.) An unusual trend is seen in the graph, maximum employees living approximately 25 kms away from work place tend to miss work more as compared to any other distance.

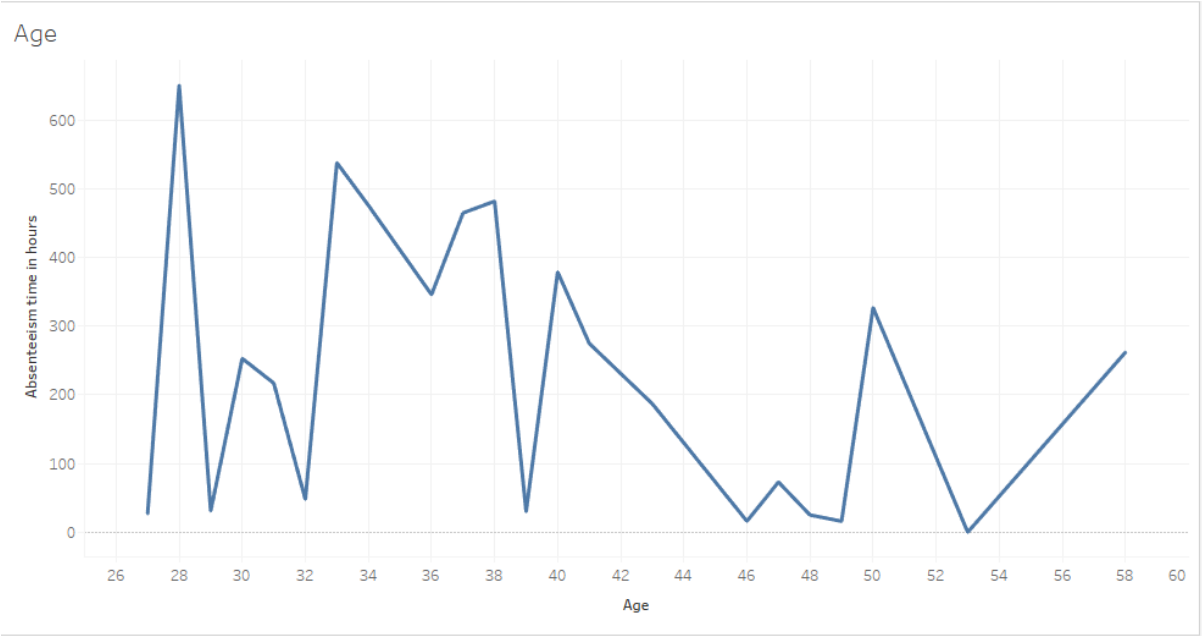
2.) Absenteeism is highest when the travel expense is between 150-200.

Dashboard 3 :

The third information dashboard is based on how age and social background of an employee affect his/her absenteeism at work

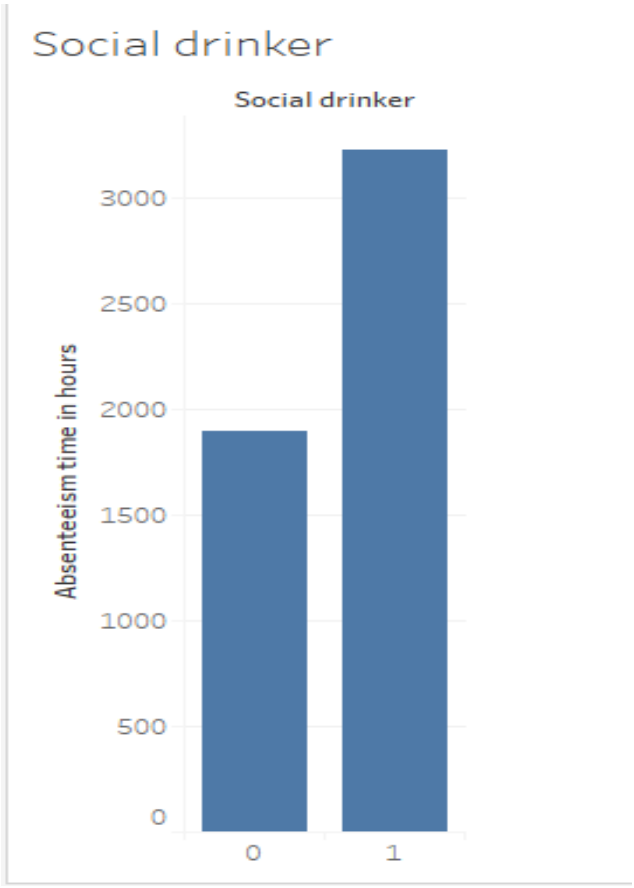
Worksheet 1 :

A graph is plotted using age as a dimension and absenteeism as a measure . This is to find out which age group tends to miss work more :



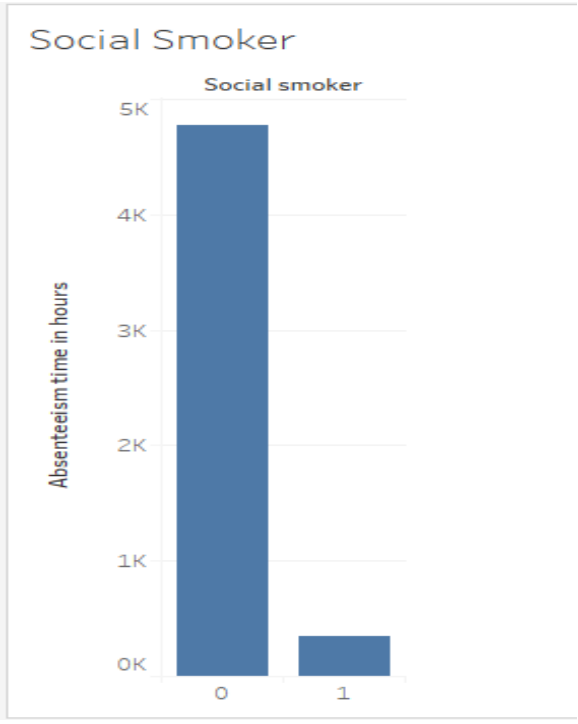
Worksheet 2 :

A graph is plotted to check how many employees are social drinkers and how it affects their absenteeism at work:

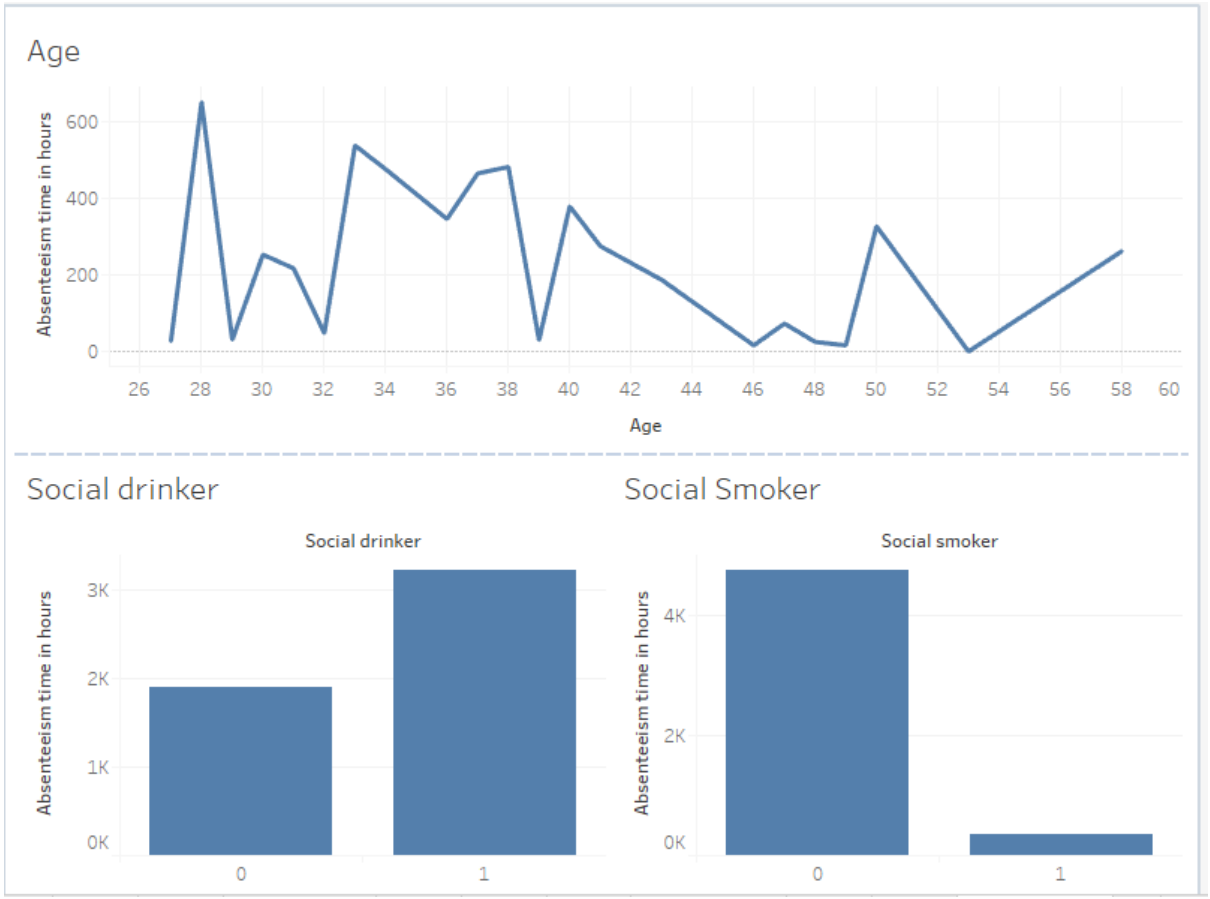


Worksheet 3 :

A graph is plotted to check how many employees are social smokers and how it affects their absenteeism at work:



The final dashboard :



Insights from the above dashboard :

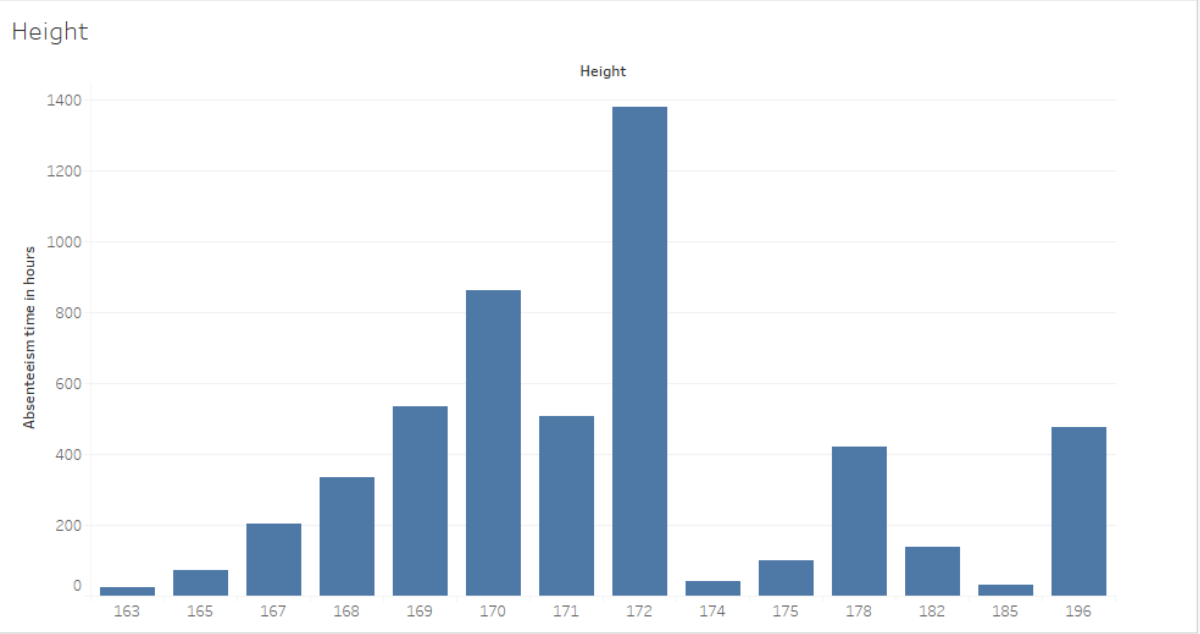
1.) The age group of 28-30 and 32-38 tend to miss work more than other age groups.
2.) Social drinkers miss work more.
3.) Unusually, non-smokers miss work more as compared to smokers.

Dashboard 4 :

The forth information dashboard is used to find out how physical fitness of an employee affects absenteeism. The features taken into consideration are height, weight and body mass index.

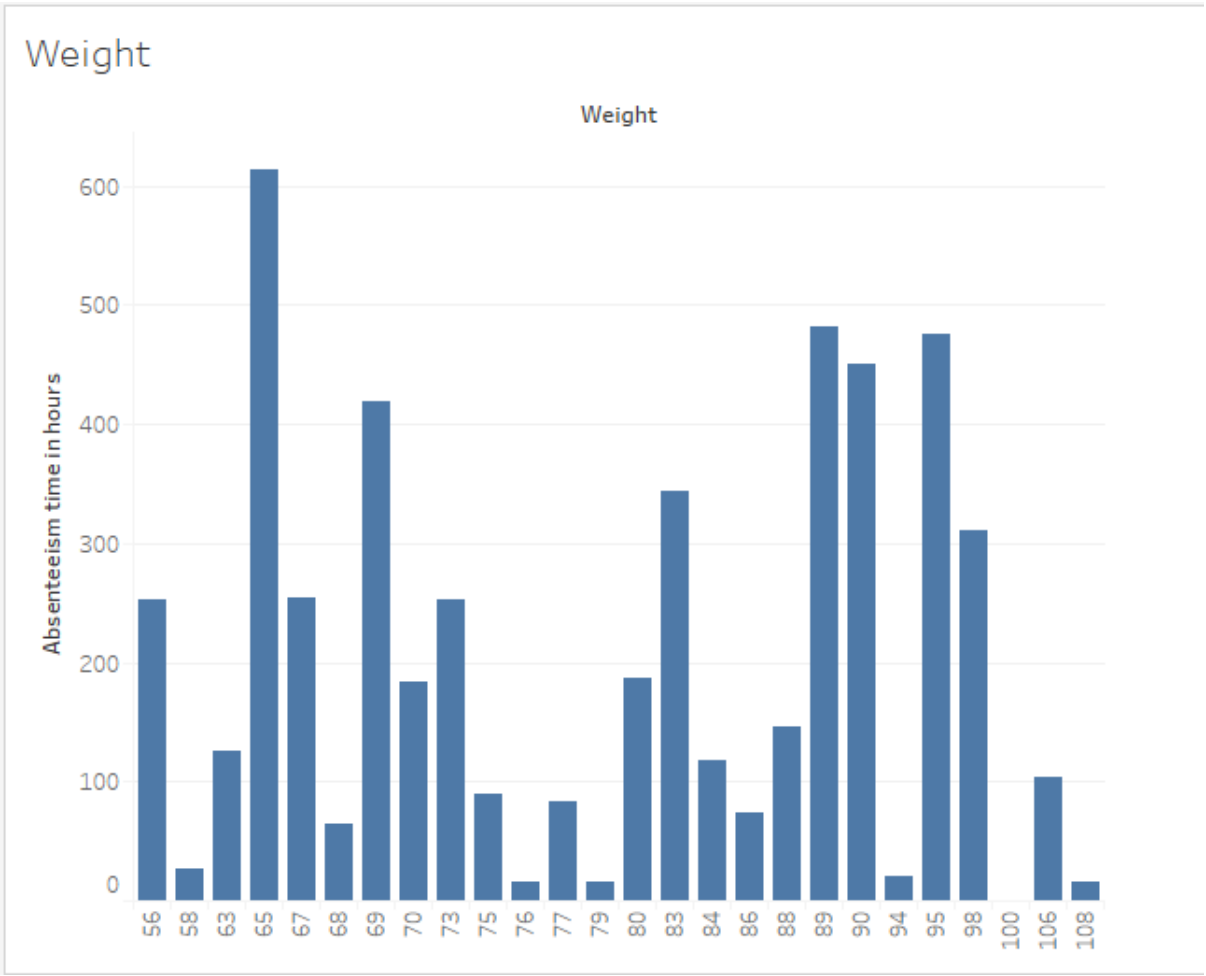
Worksheet 1 :

A graph is plotted using height as the dimension and absenteeism in hours as the measure :



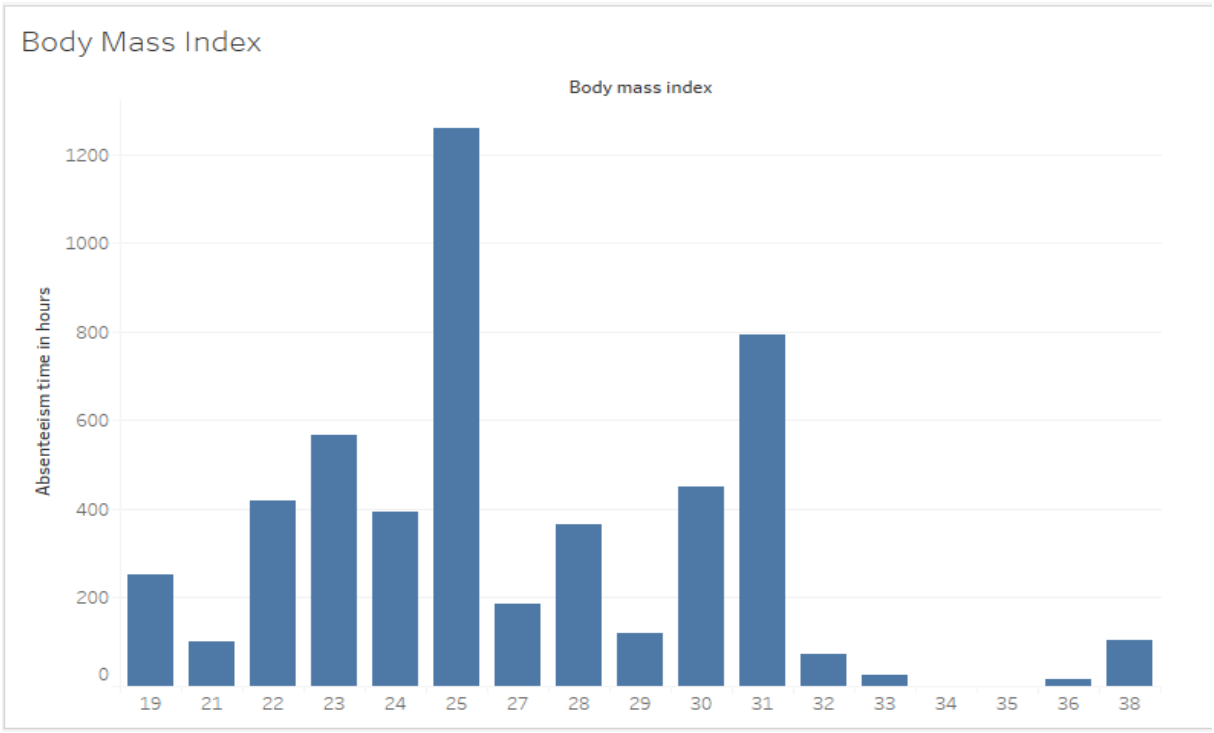
Worksheet 2 :

A graph is plotted using weight as the dimension and absenteeism in hours as the measure :

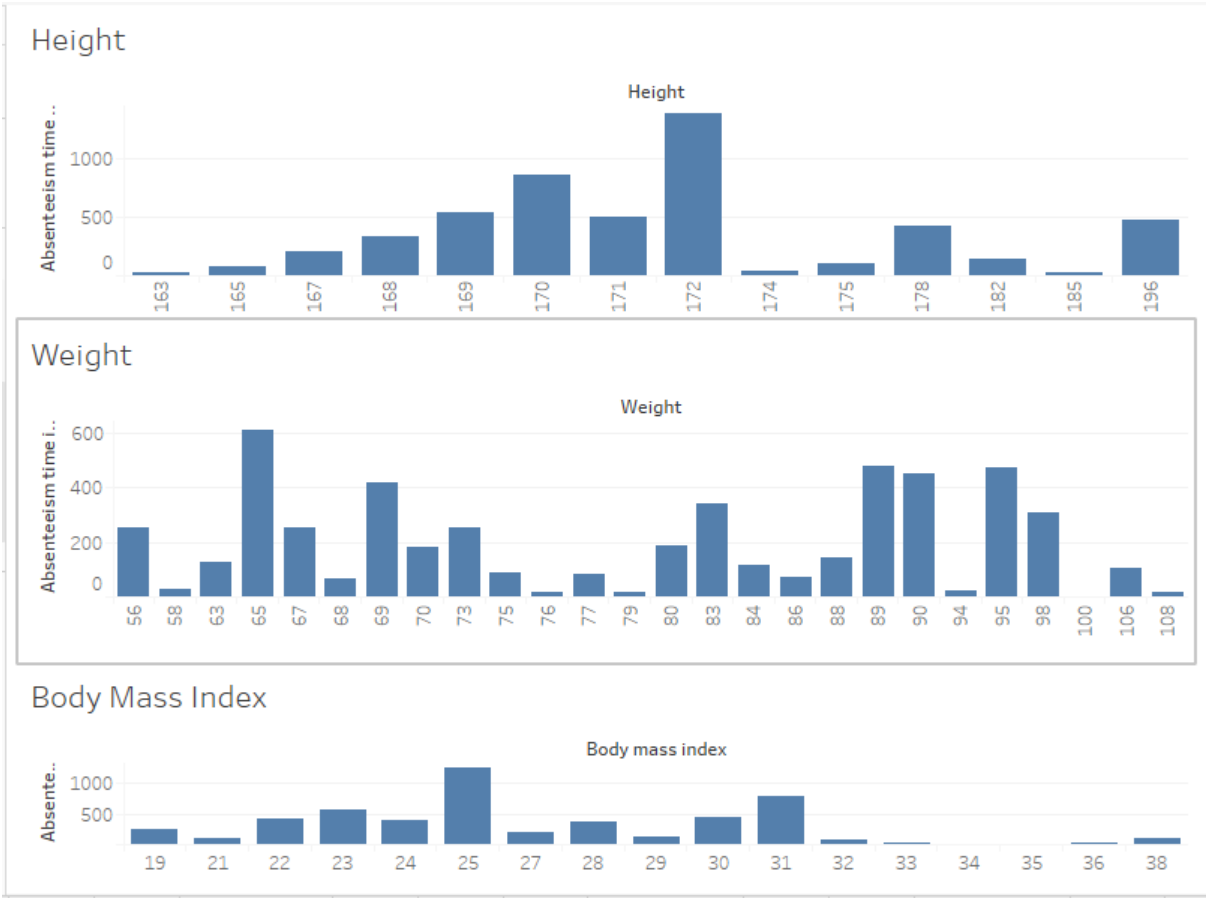


Worksheet 1 :

A graph is plotted using body mass index as the dimension and absenteeism in hours as the measure :



The final dashboard :



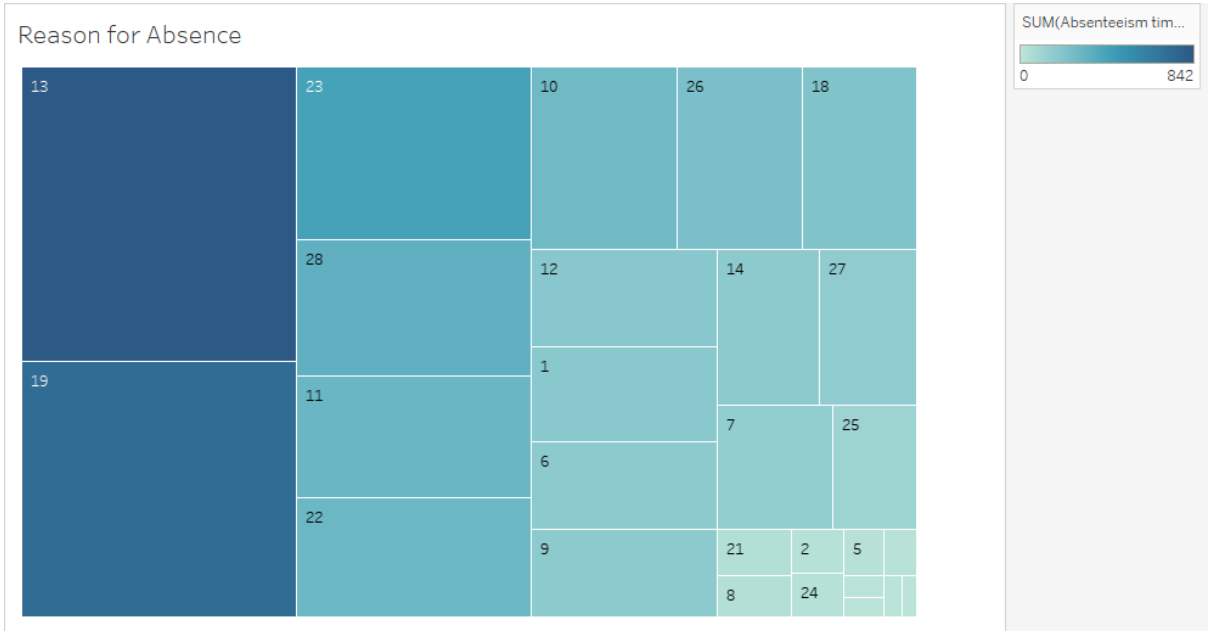
Insights from the dashboard :

1. Employees with an average height of 172 tend to miss work more .(this can be irrelevant)

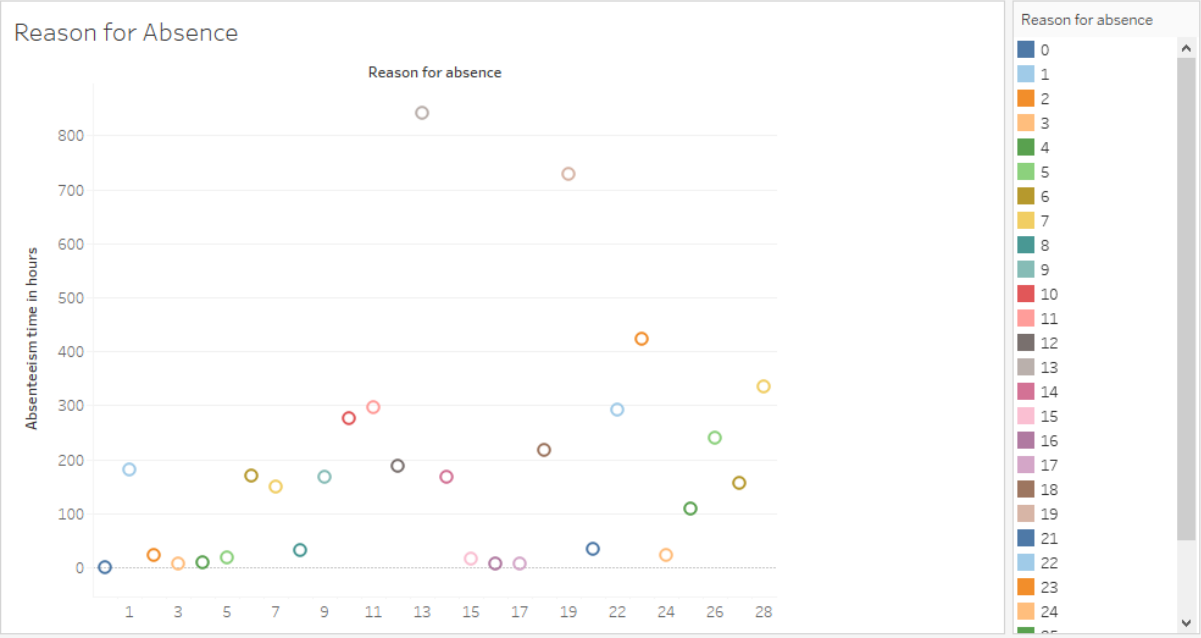
2. It is difficult to predict how weight affects target, as it is randomly spread out for different types of weight.
3. Employees with a low BMI tend to miss work more.

Dashboard 5 :
The fifth information dashboard gives information about which reason of absence is given the most for absenteeism. We use two different graphs for checking the influence. One is a tree map and second one is the scatter plot graph. In both the graphs reason for absence is used as a dimension and absenteeism in hours is used as a measure.

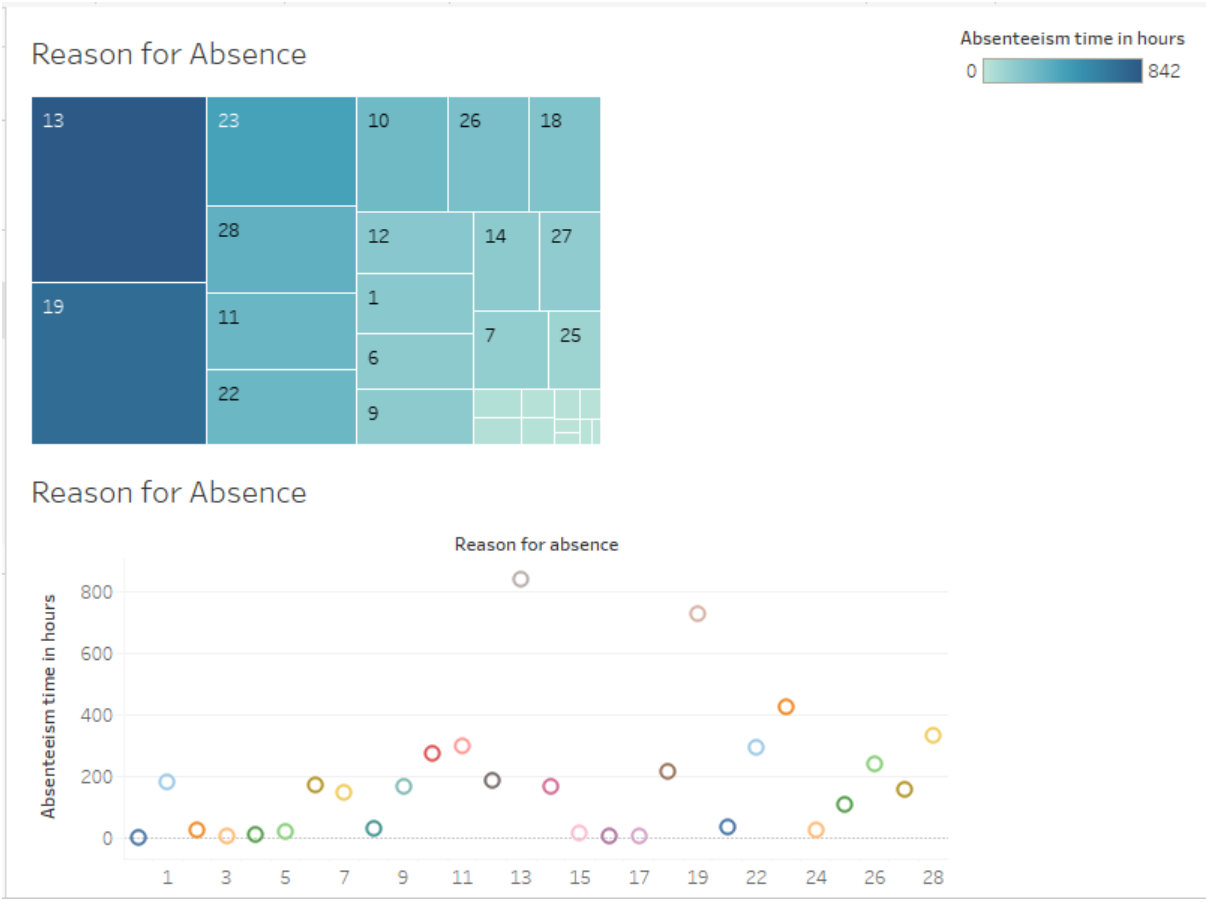
Tree map :



Scatter plot :



The final dashboard :



Insights from the dashboard :

1.) As we can see, in both the graphs reason number 13 corresponds to more absenteeism followed by reason 19.
2.) Hence health problems like connective tissue injury, infections , food poisoning affect absenteeism on a large scale .

Dashboard 6 :

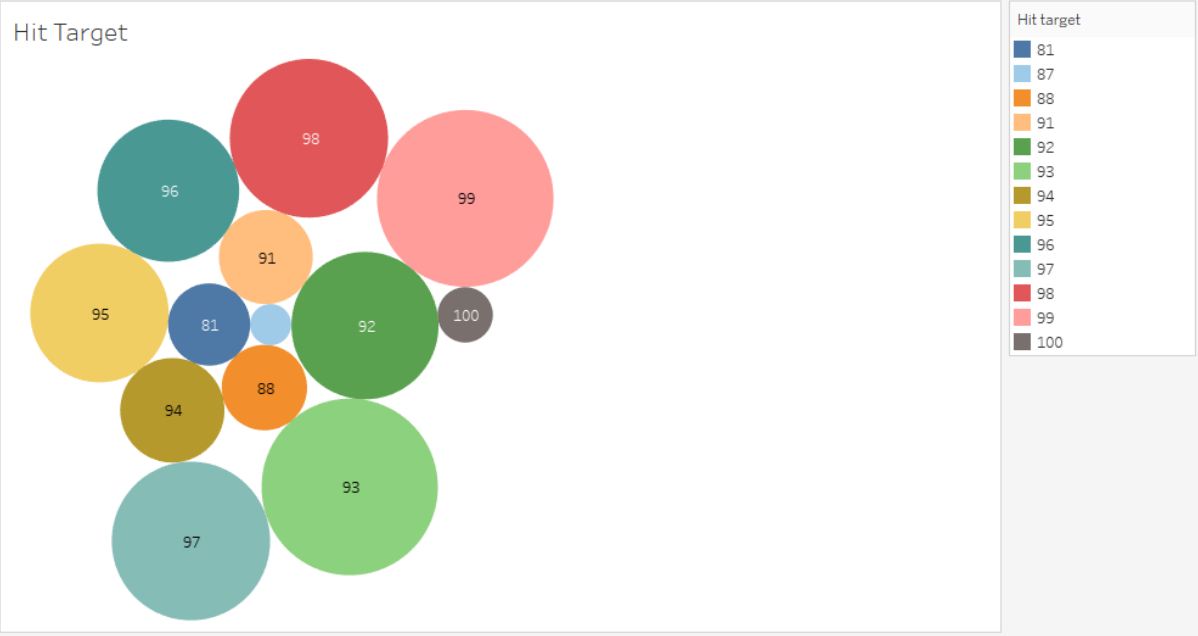
The sixth information dashboard provides information about work pressure and performance affect absenteeism at work. The attributes taken into consideration while developing this dashboard are as follows :

1. Hit target in percentage
2. Work load average/day
3. Service time

Different types of dashboards are developed, and each of the above attribute is mapped against absenteeism in hours as the measure :

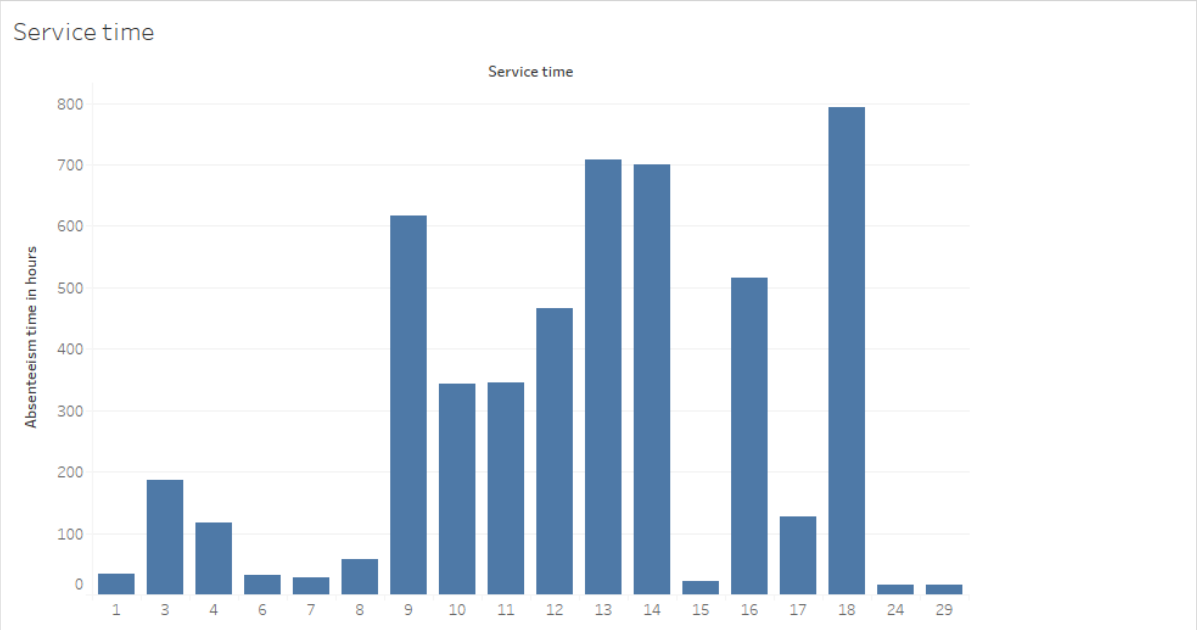
Worksheet 1 :

Using the Hit target , bubble chart is plotted . This gives us an idea what is the relationship between achieving a target given to absenteeism at work .



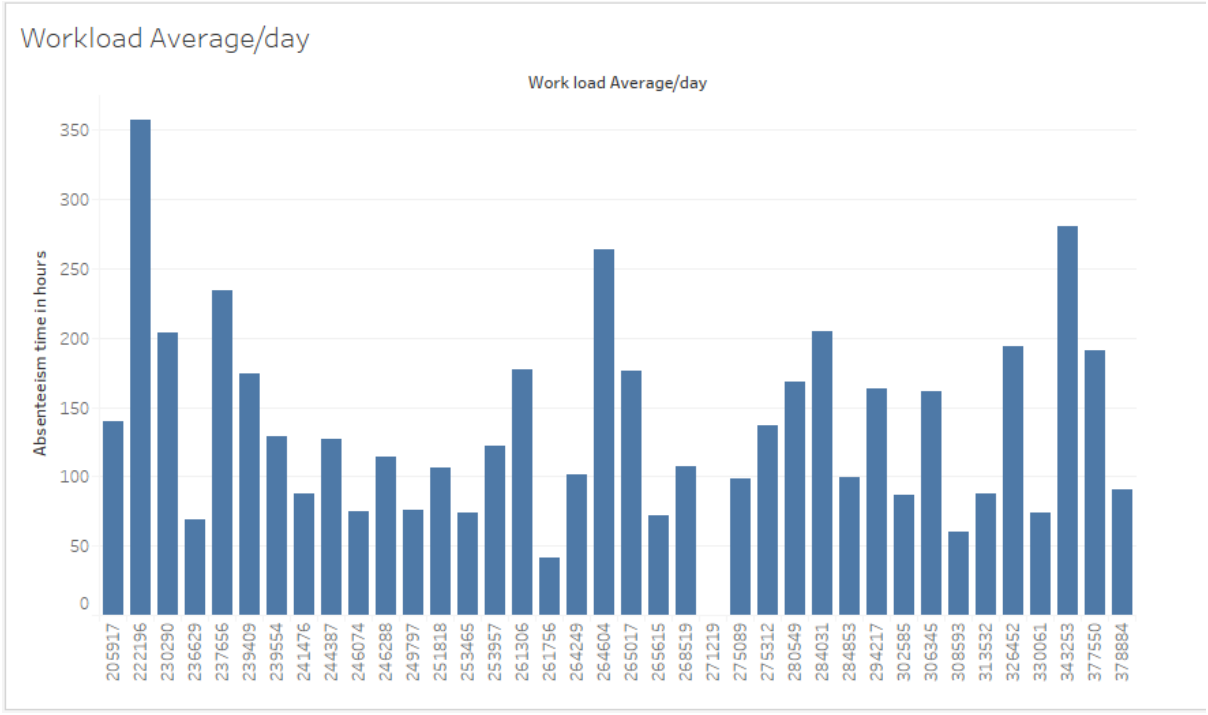
Worksheet 2 :

Using service time as a measure a bar chart is plotted. This will help us to find out if this attribute is related to absenteeism in hours linearly or inversely or any other kind of relationship.

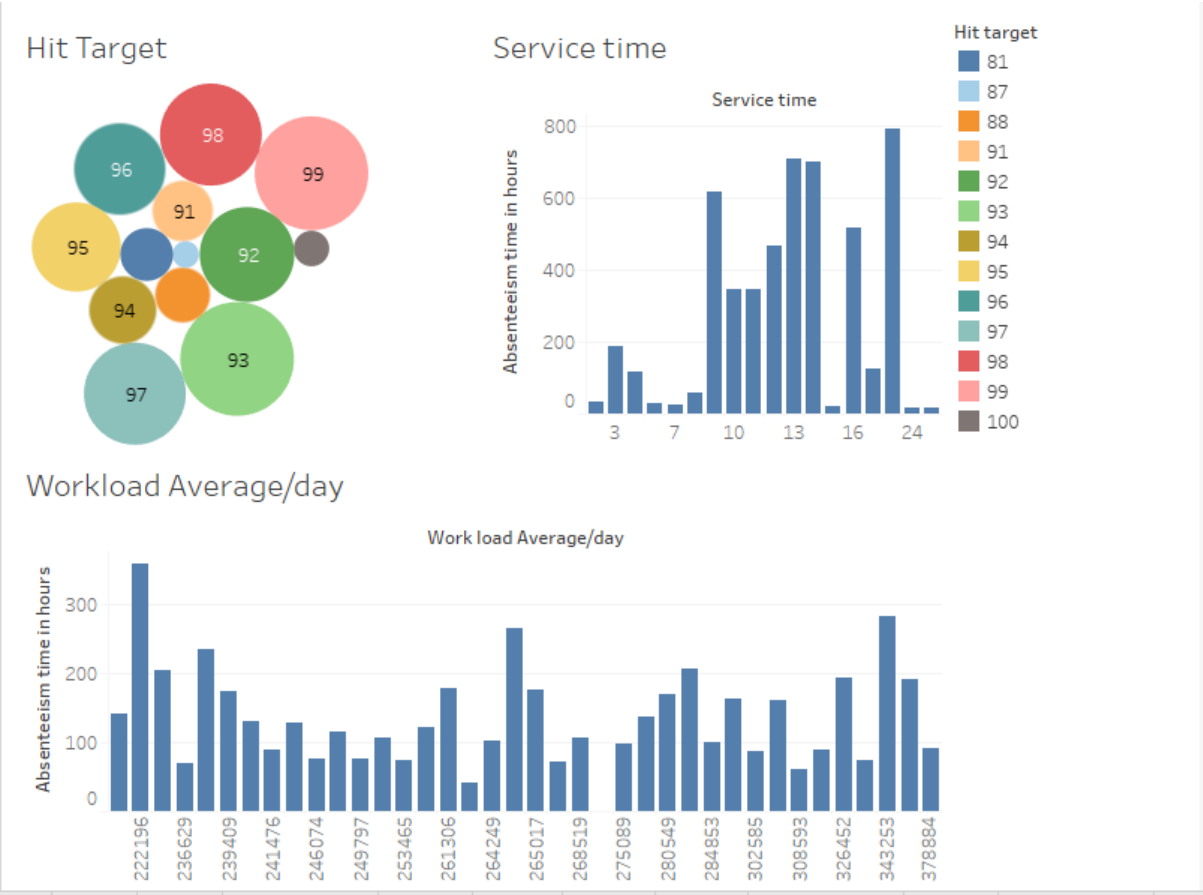


Worksheet 3 :

In this graph work load average per day is used as a dimension. This can be considered as an important attribute .



The final dashboard :



Insights from the dashboard :

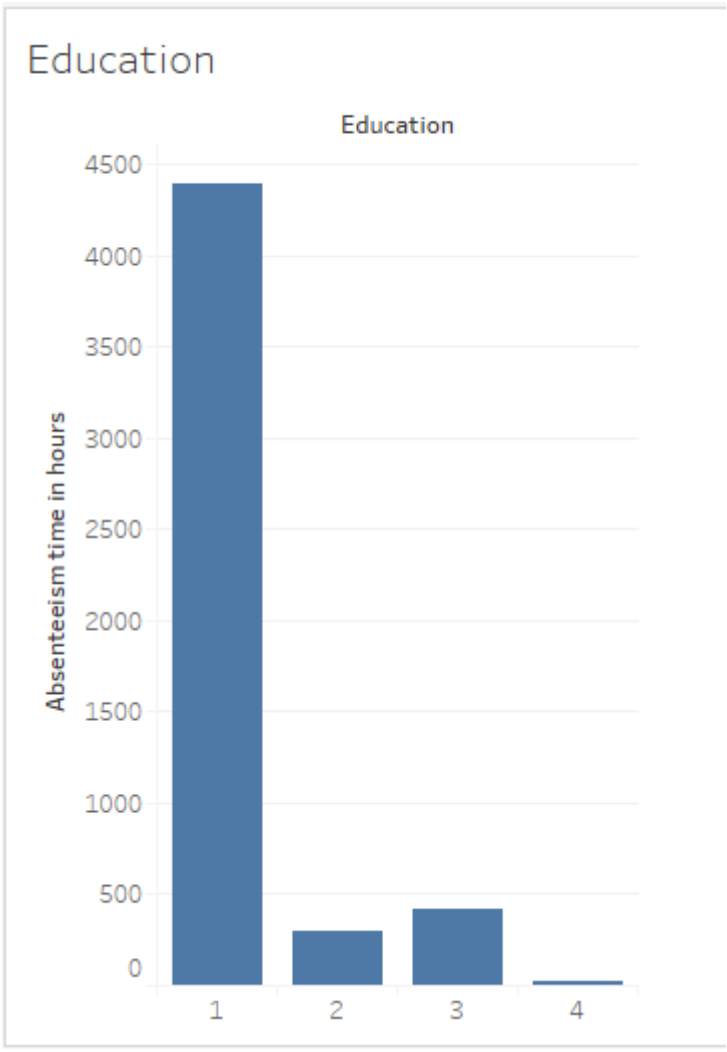
1.) Employees that have achieved 99 percent of target to miss work more
2.) Absenteeism is more when the working hours lie in the range of 13 to 18.

Dashboard 7 :

In this dashboard we check how attribute education and age affect the absenteeism at work.

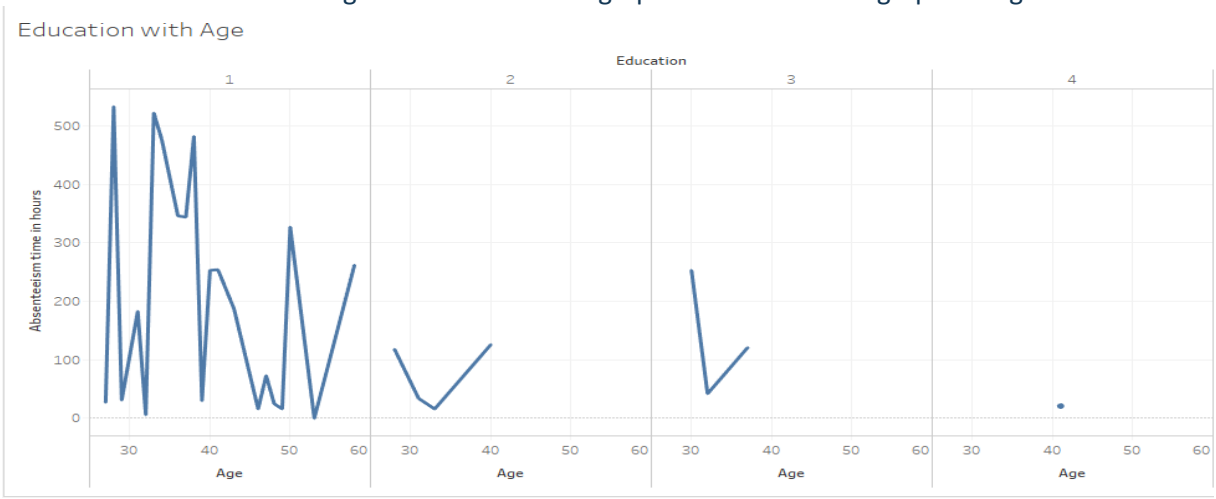
Work sheet 1 :

Initially we just plot education against absenteeism and develop a bar chart .



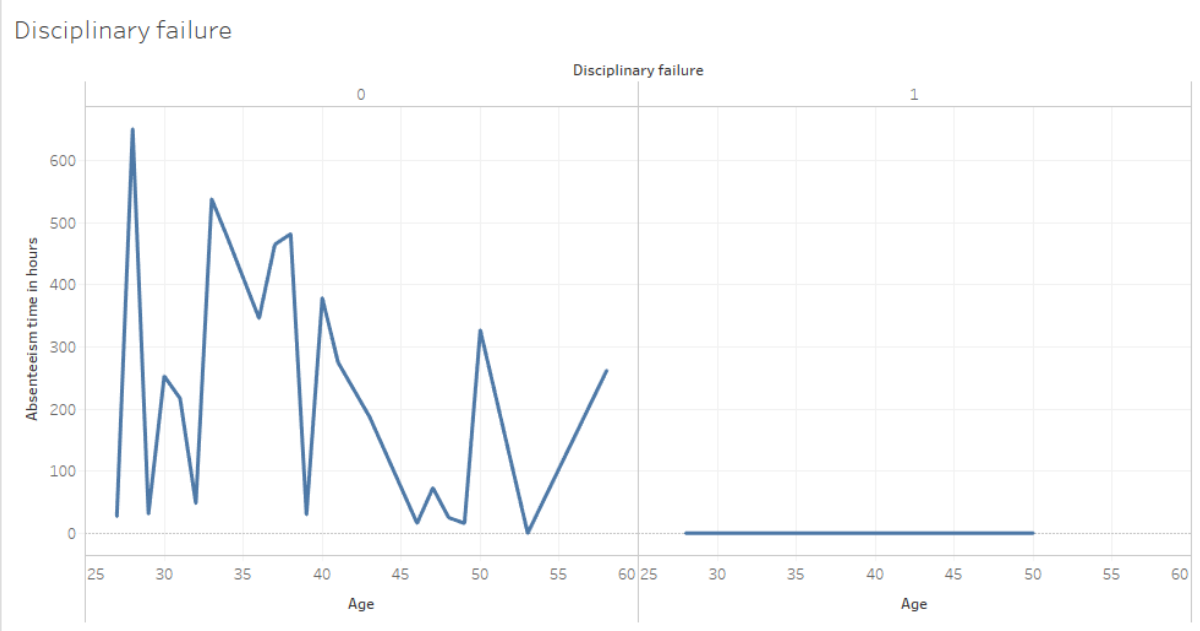
Worksheet 2 :

Now we also introduce the age attribute in above graph and see how the graph changes:

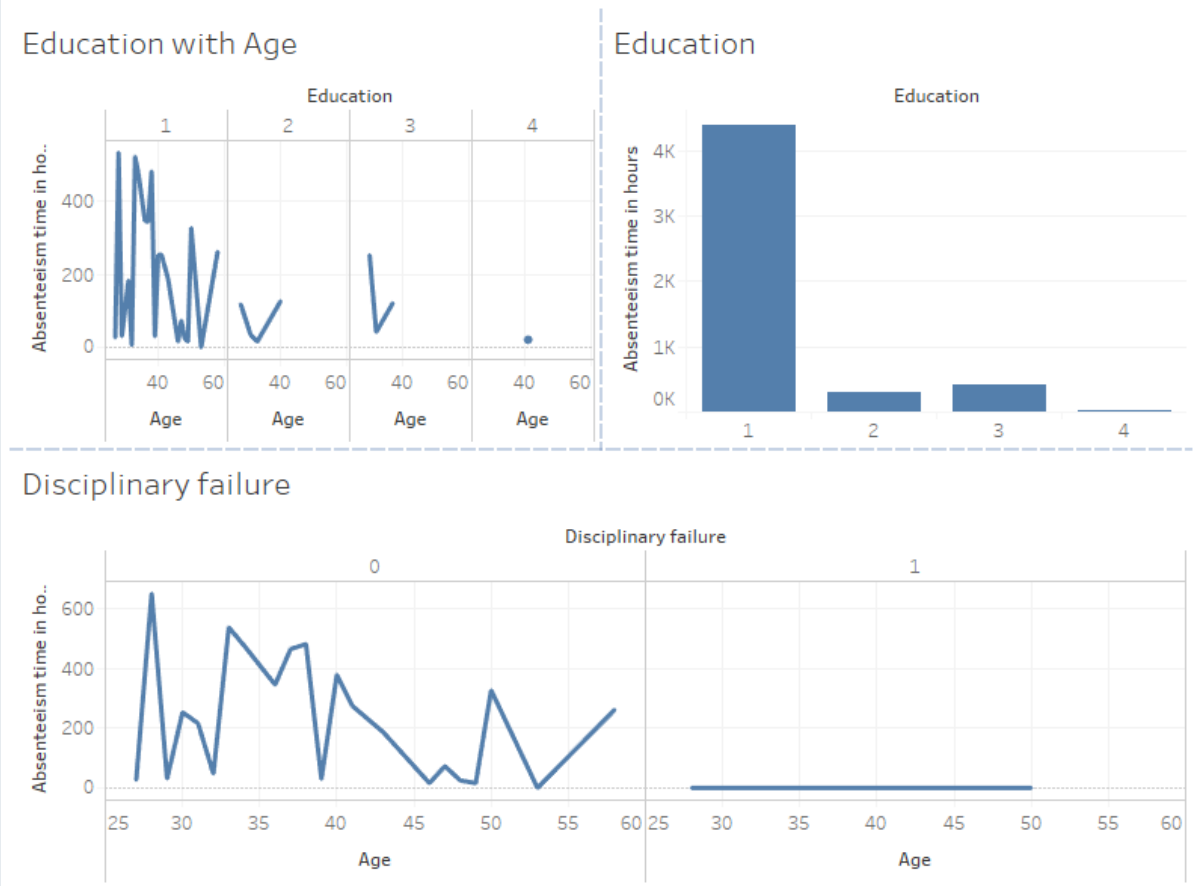


Worksheet 3 :

Now we take consideration the age attribute and the disciplinary failure attribute and check how this affects the absenteeism at work.



The final dashboard :



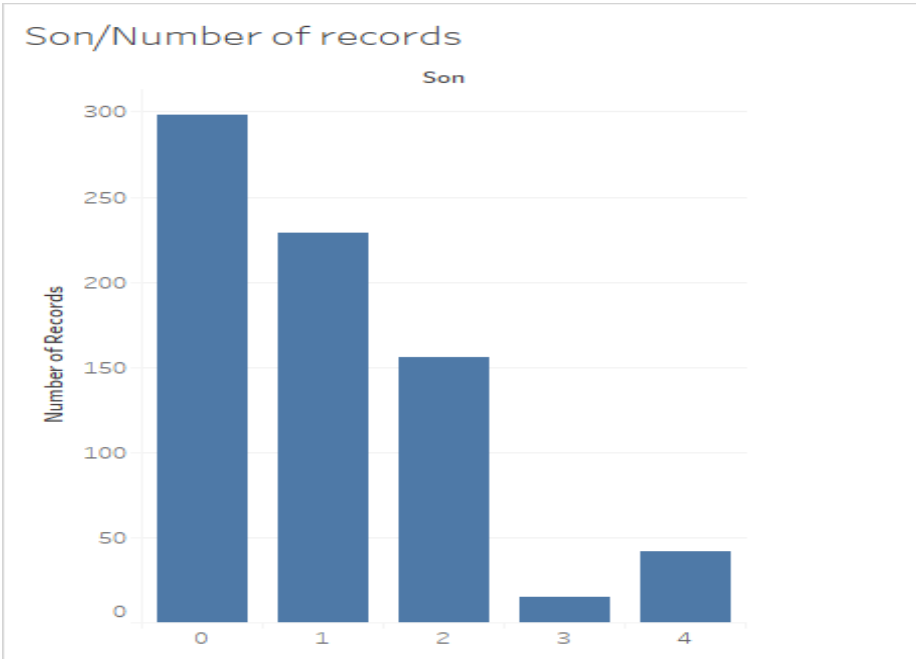
Insights from the dashboard :

1) People with just high school level of education are tend to miss work more.
2) The age group 0f 25-35 tends to miss work more
3) Chances of not being a disciplinary failure and missing work are negligible .

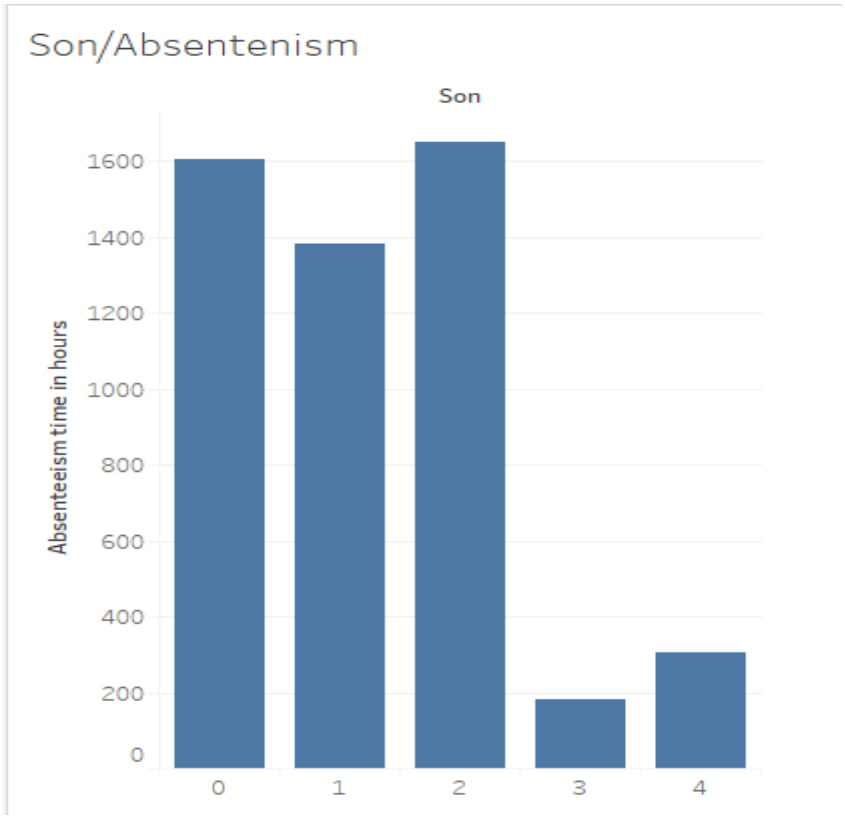
Dashboard 8 :

In this dashboard , we check how the family life of an employee affects the target variable by taking into consideration attributes such as number of children and number of pets the employee has. The above two attributes are selected as dimensions and absenteeism is selected as a measure.

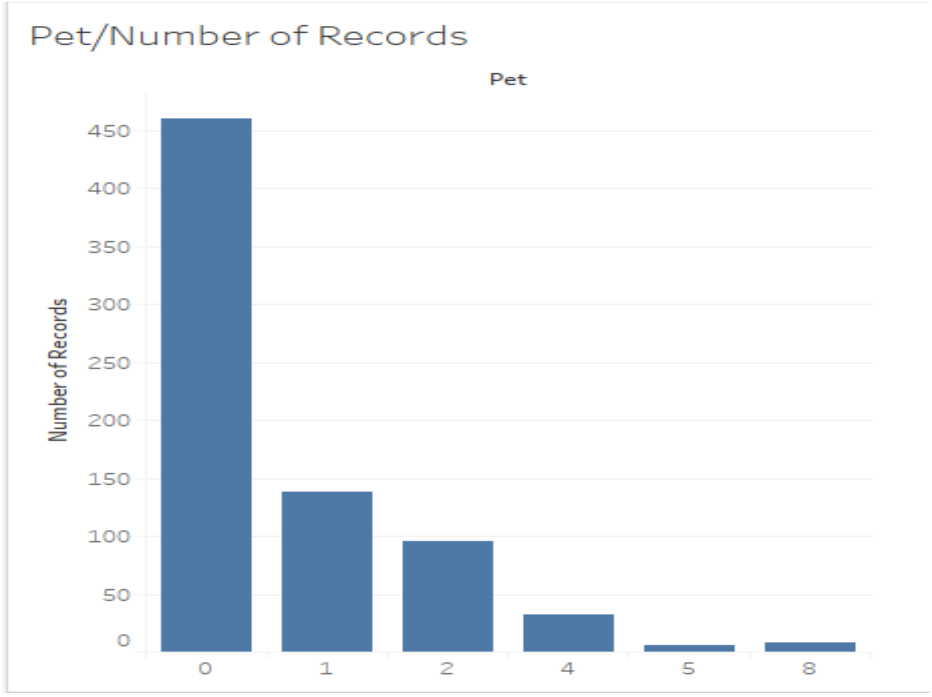
Worksheet 1 :



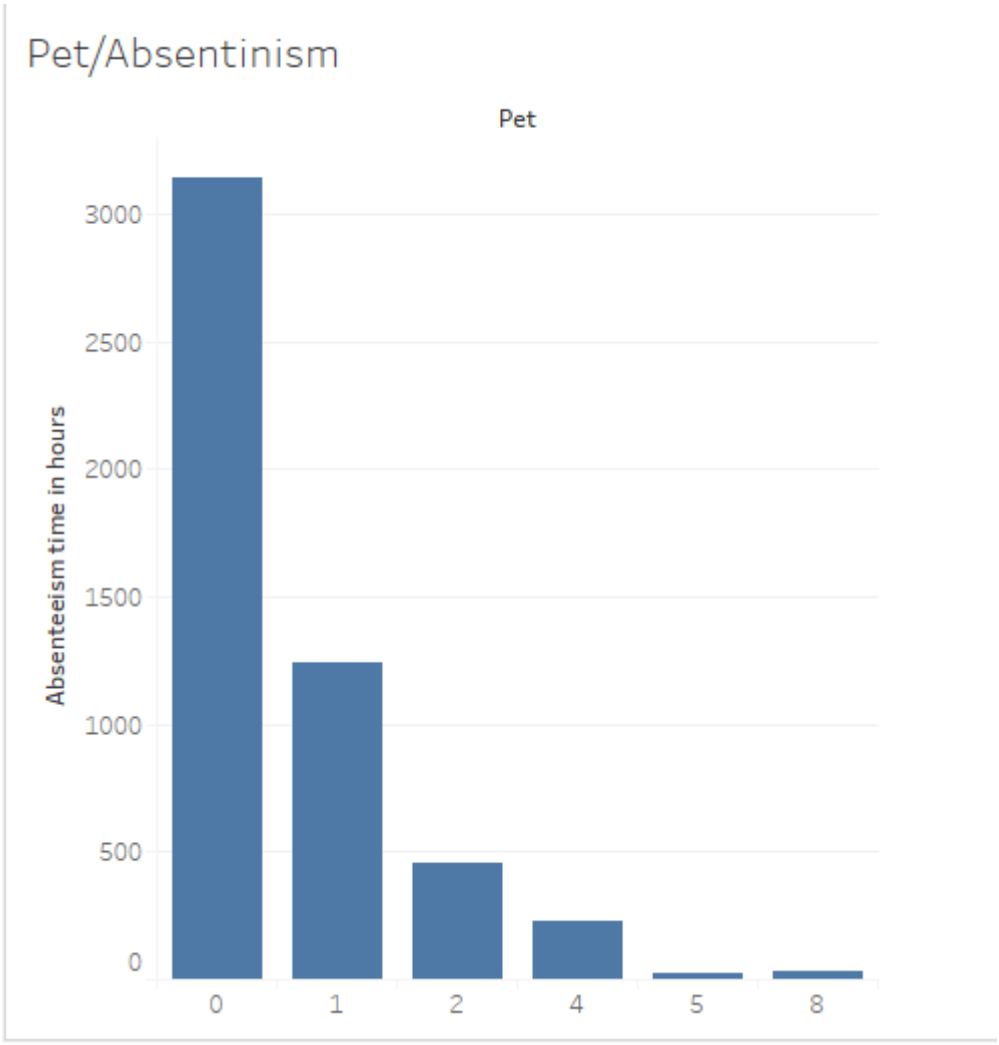
Worksheet 2 :



Worksheet 3 :



Worksheet 4 :



The final dashboard :

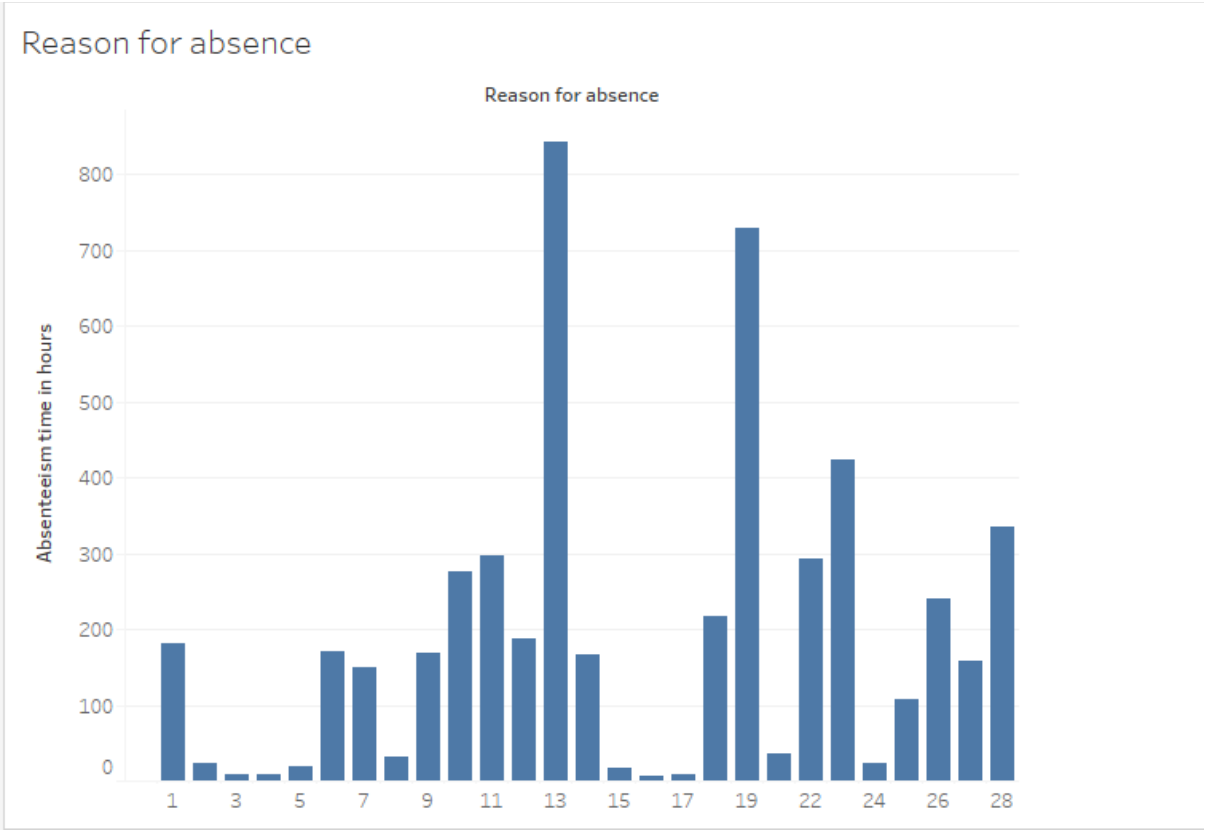
Insights from the dashboard :

1.) Employees with 2 children tend to miss work more.
2.) Employees with a 1 pet tend to miss work more.

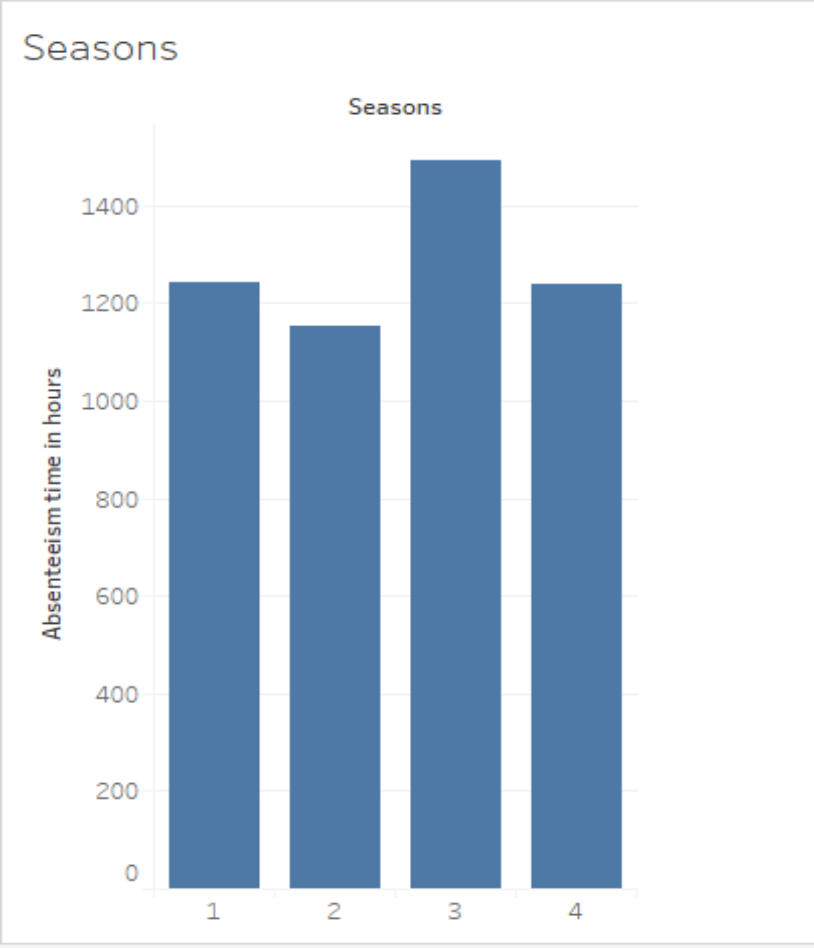
Dashboard 9 :

In this dashboard we try to find out if reasons for absence i.e. sickness and seasons such as summer, rainy etc have any relationship between them and how do they together affect absenteeism at work .

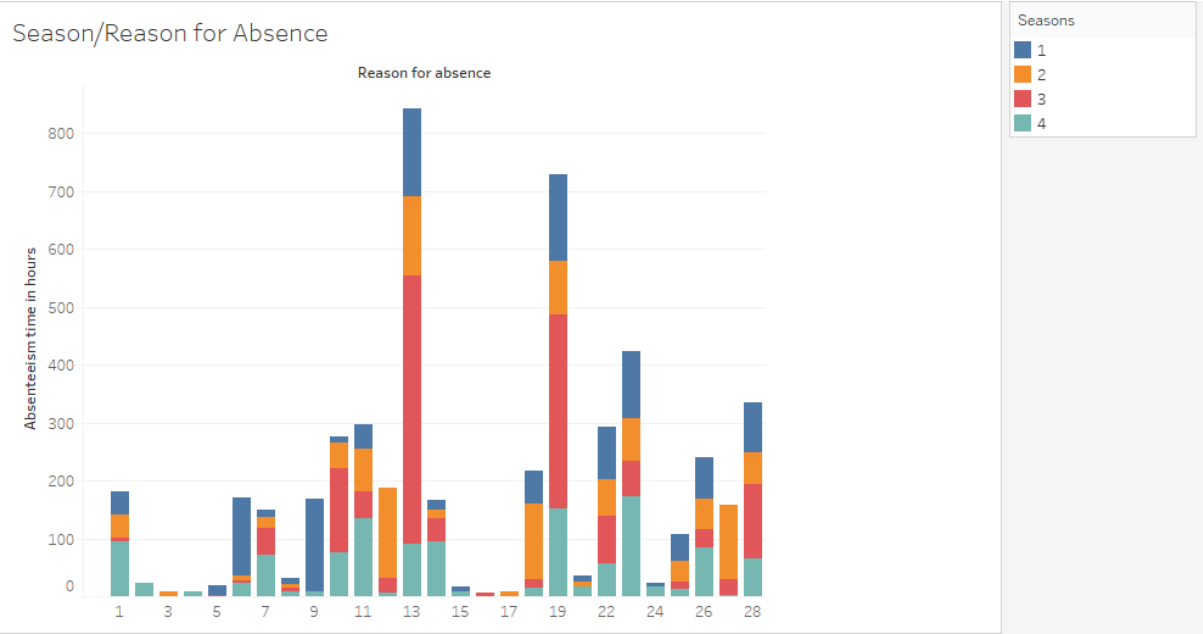
Worksheet 1 :


Reason for absence
Reason for absence
Absenteeism time in hours
800
700
600
500
400
300
200
100
0
1
3
5
7
9
11
13
15
17
19
22
24
26
28

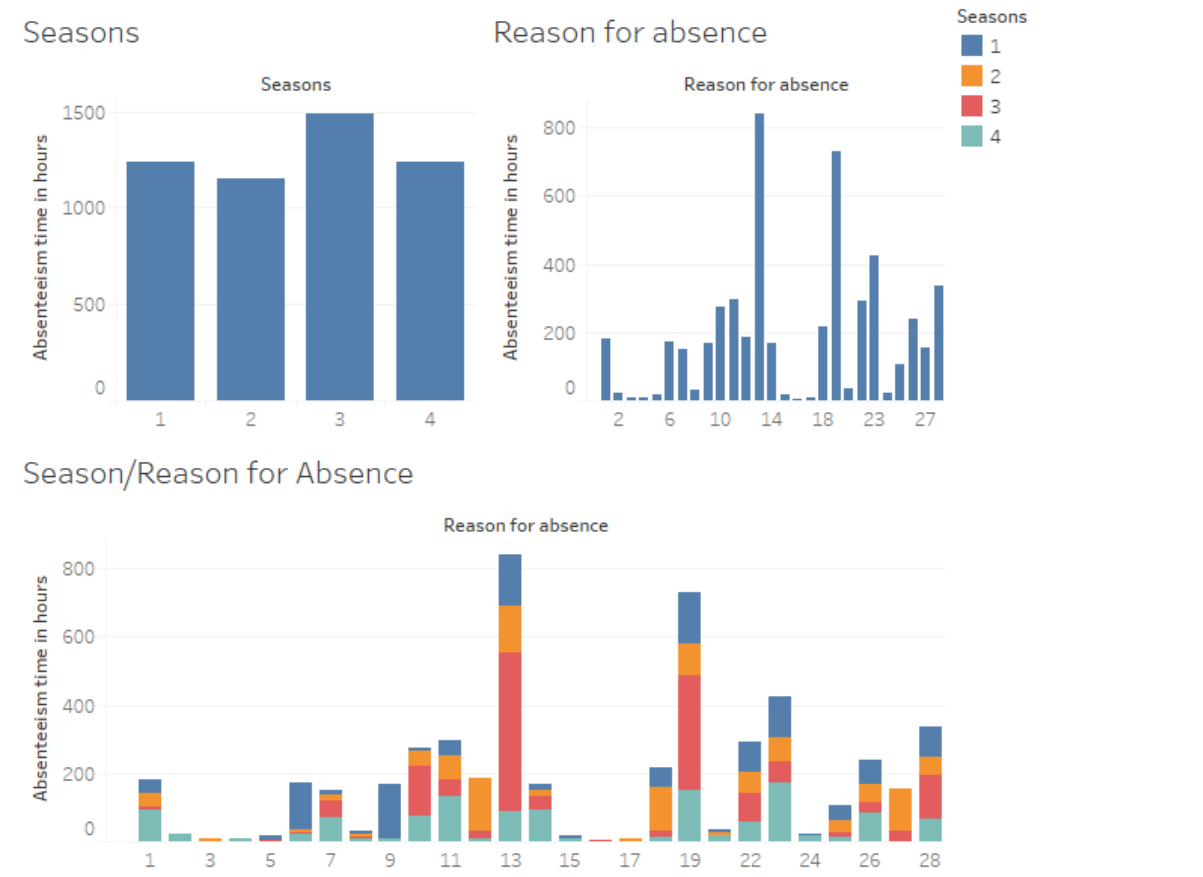
Worksheet 2 :


Seasons
Seasons
Absenteeism time in hours
1400
1200
1000
800
600
400
200
0
1
2
3
4

Worksheet 3 :



The final dashboard :



Insights :

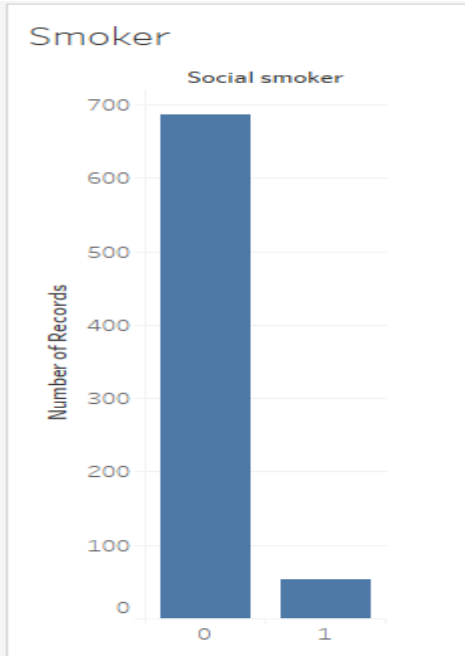
1.)Season 3 has more absenteeism than any other season.

2.) Reason 13 and 19 remain the highest even if the season attribute is introduced.

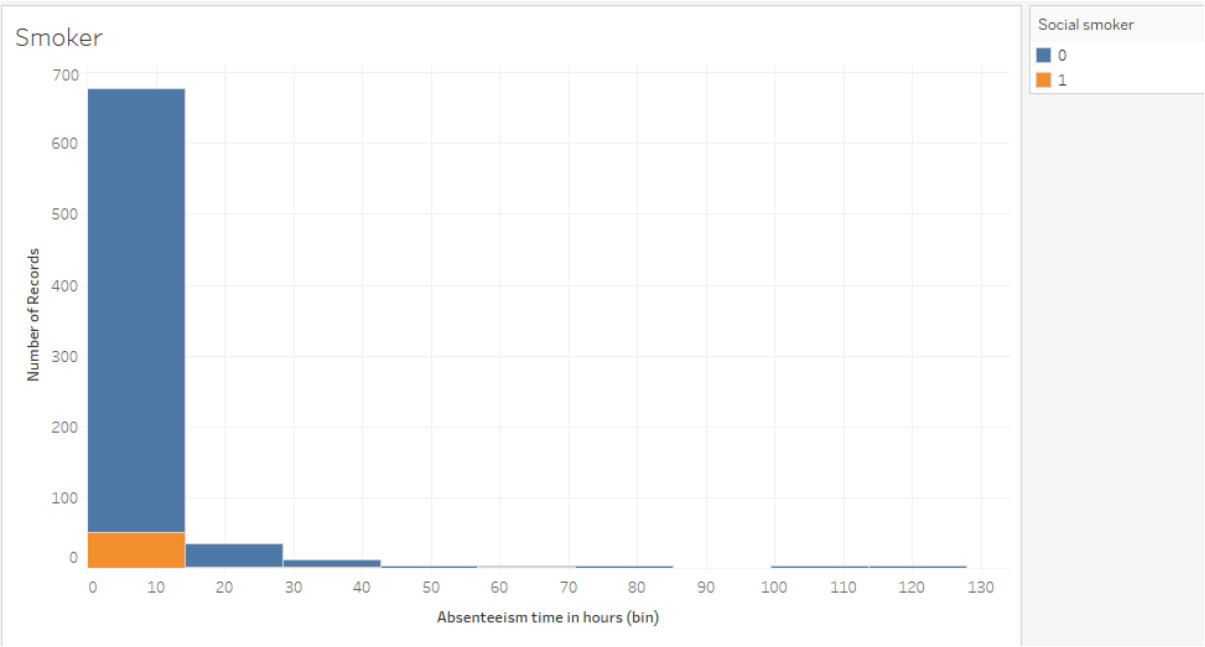
Dashboard 10 :

In this dashboard , we see how the drinking as smoking affects the absenteeism . We also use the number of records as an attribute for more detailed results :

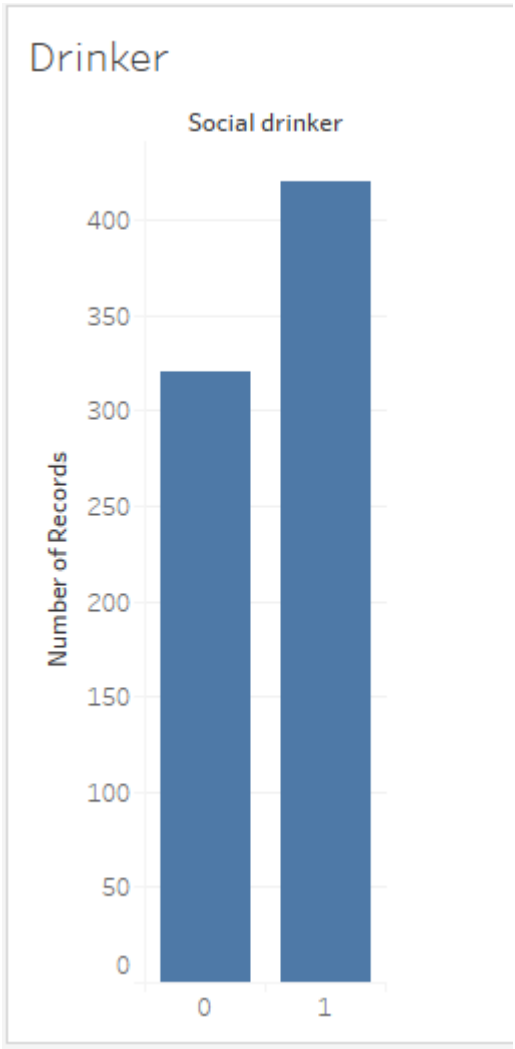
Worksheet 1 :



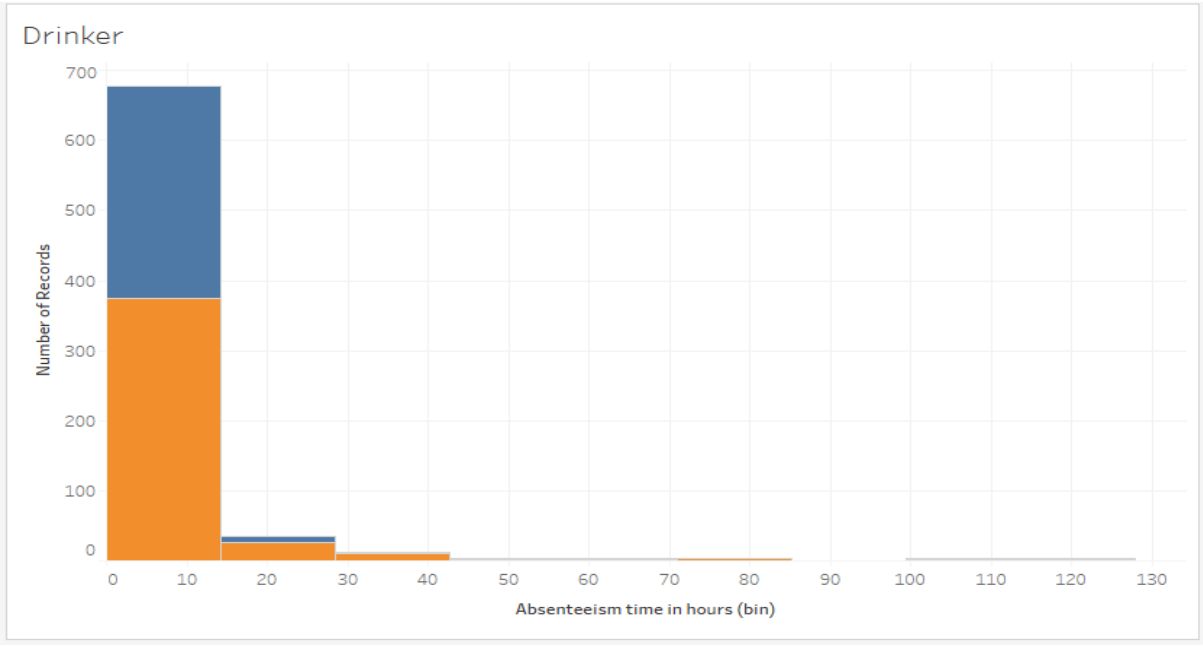
Worksheet 2 :



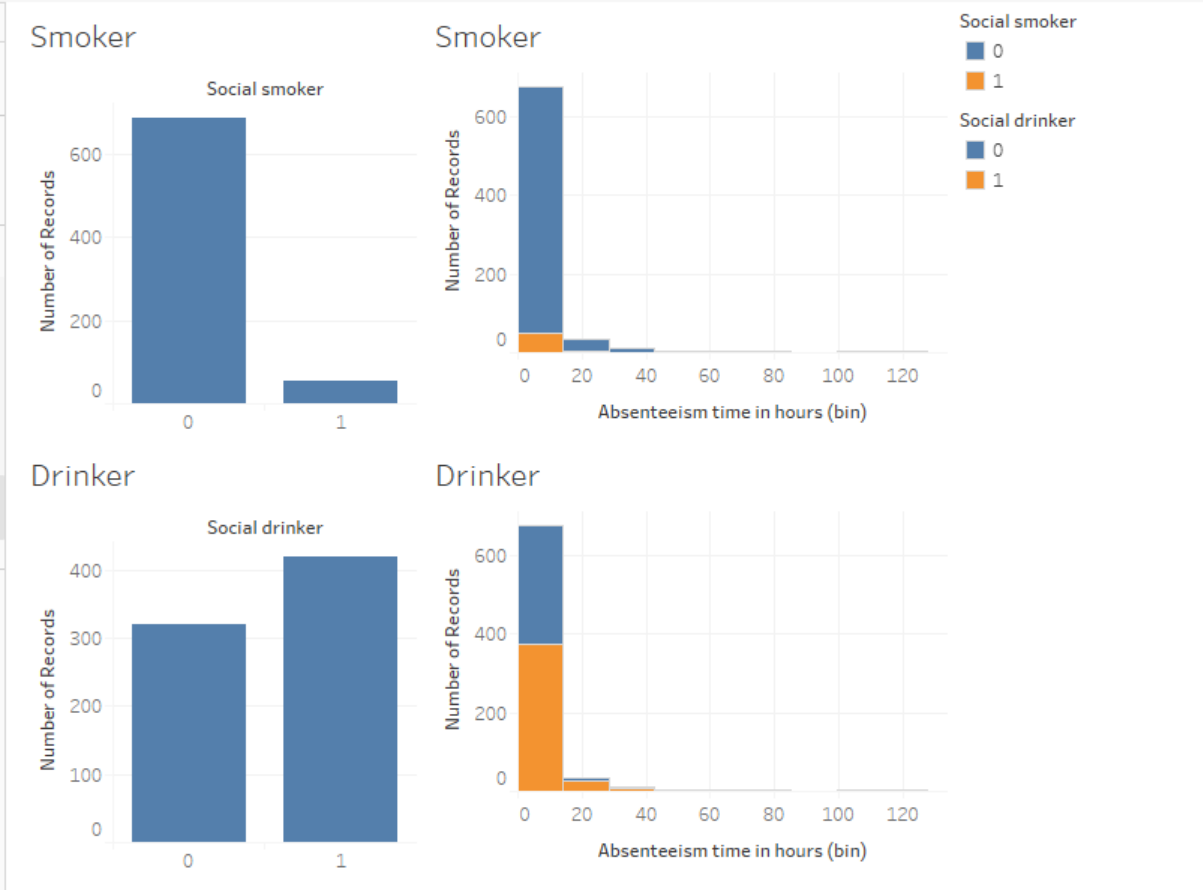
Worksheet 3 :



Worksheet 4 :



The final dashboard :



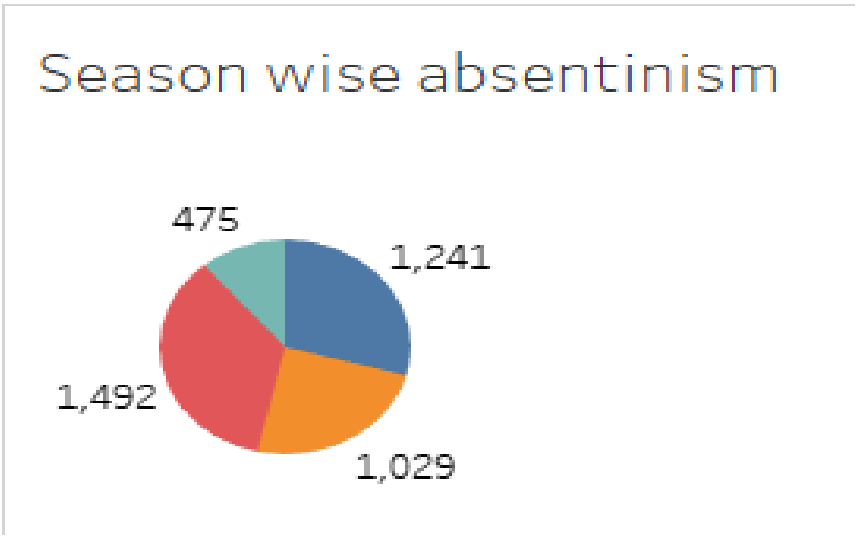
Insights :

1) Drinking affects more than smoking to absenteeism at work.

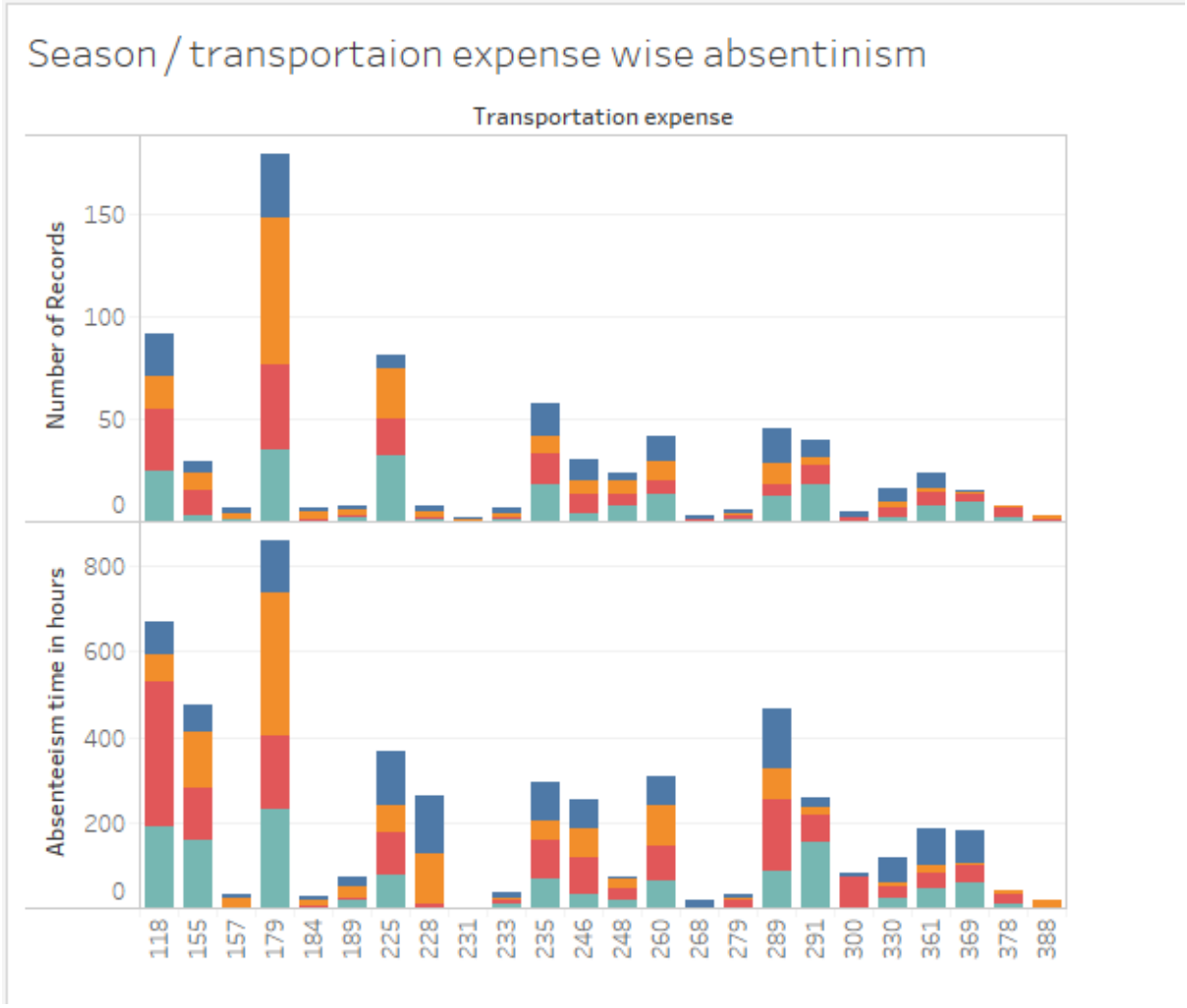
Dashboard 11 :

This dashboard gives information about season ad reason wise absenteeism ate work.

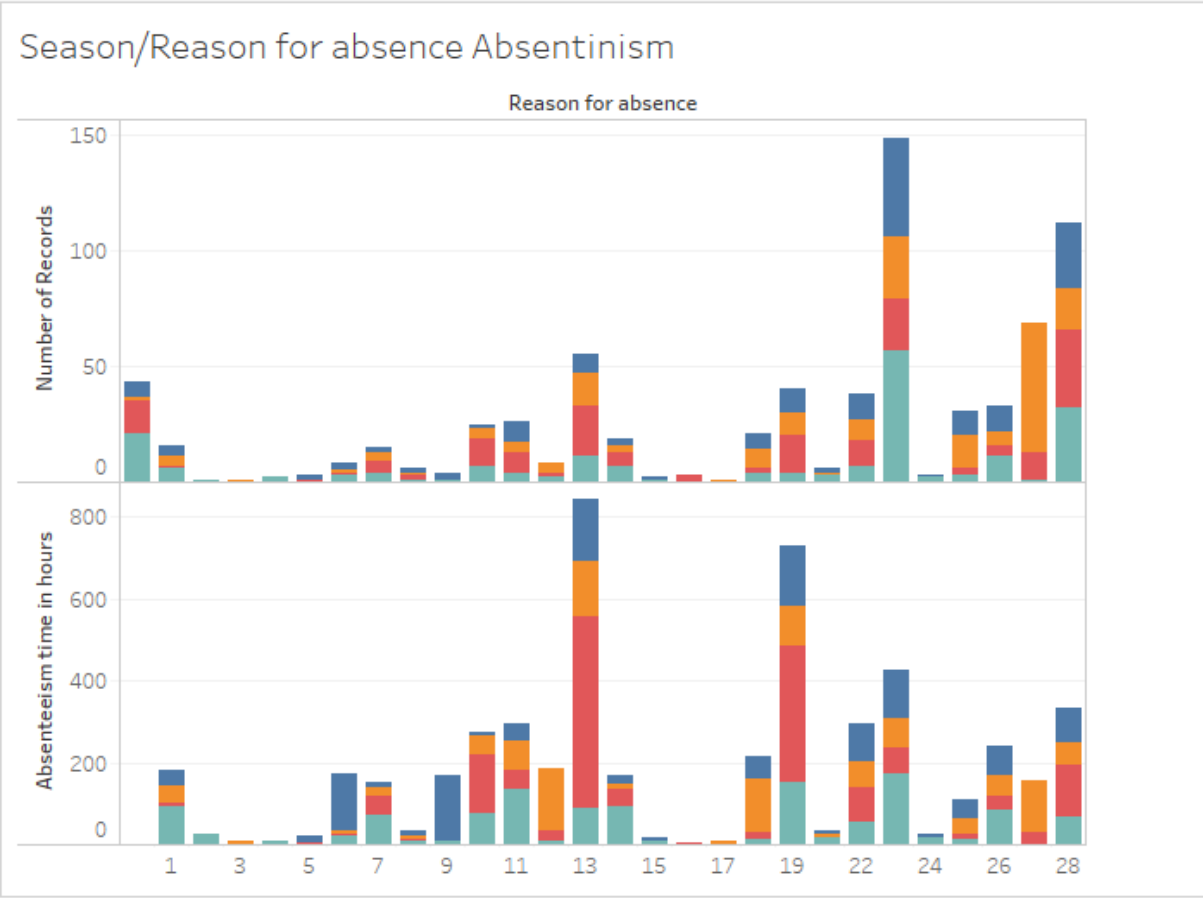
Worksheet 1 :



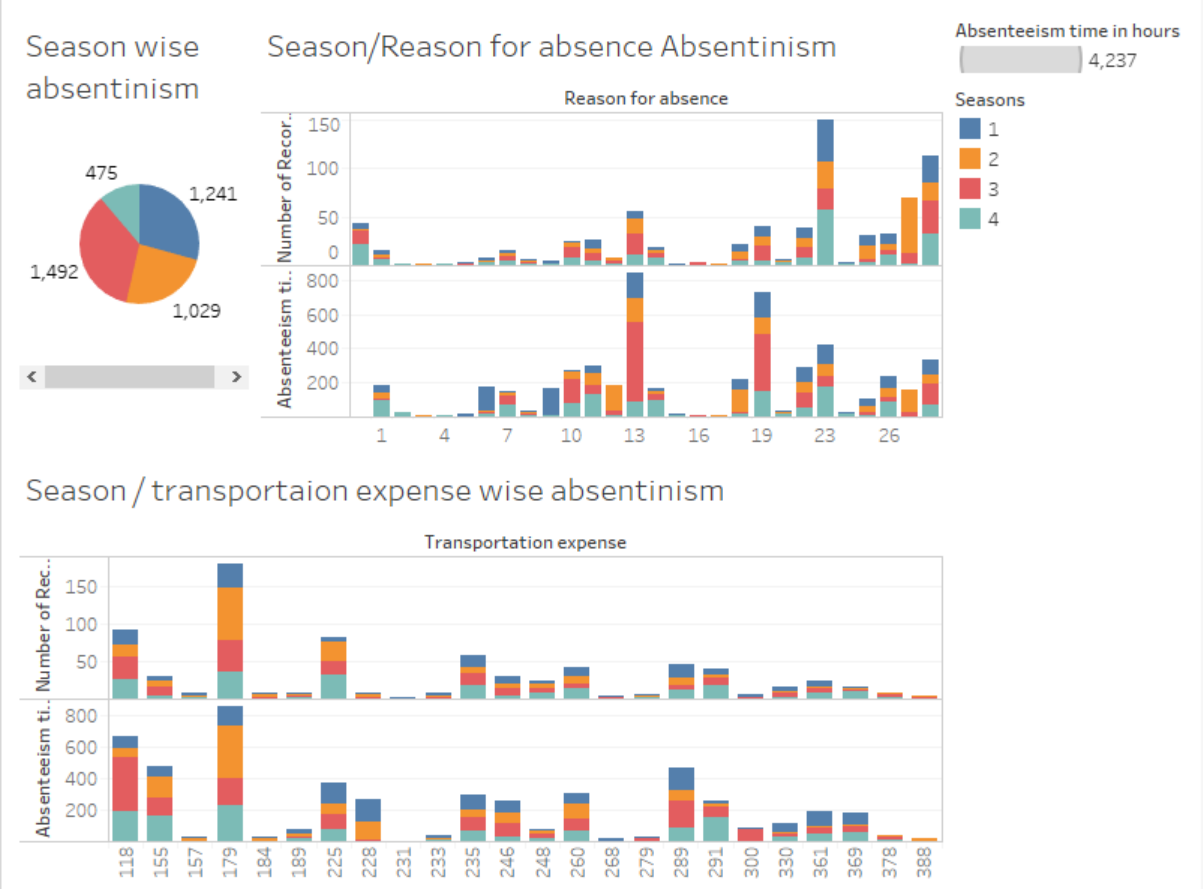
Worksheet 2 :



Worksheet 3 :



The final dashboard :



Insights :

1.) Transportation expense is related to season which is turn affects absenteeism at work
2.) The same is applicable for reason for absences.

Actionable intelligence based on insights gained from exploratory data analysis :

1) Provide transport facilities to employees that live far away to reduce absenteeism.
2) Find measure to reduce workload average per day.
3) Find if any communicable disease is affecting employees and hence take medical actions to prevent the above.
4) Design work schedule according to seasons, i.e. depending upon the weather.

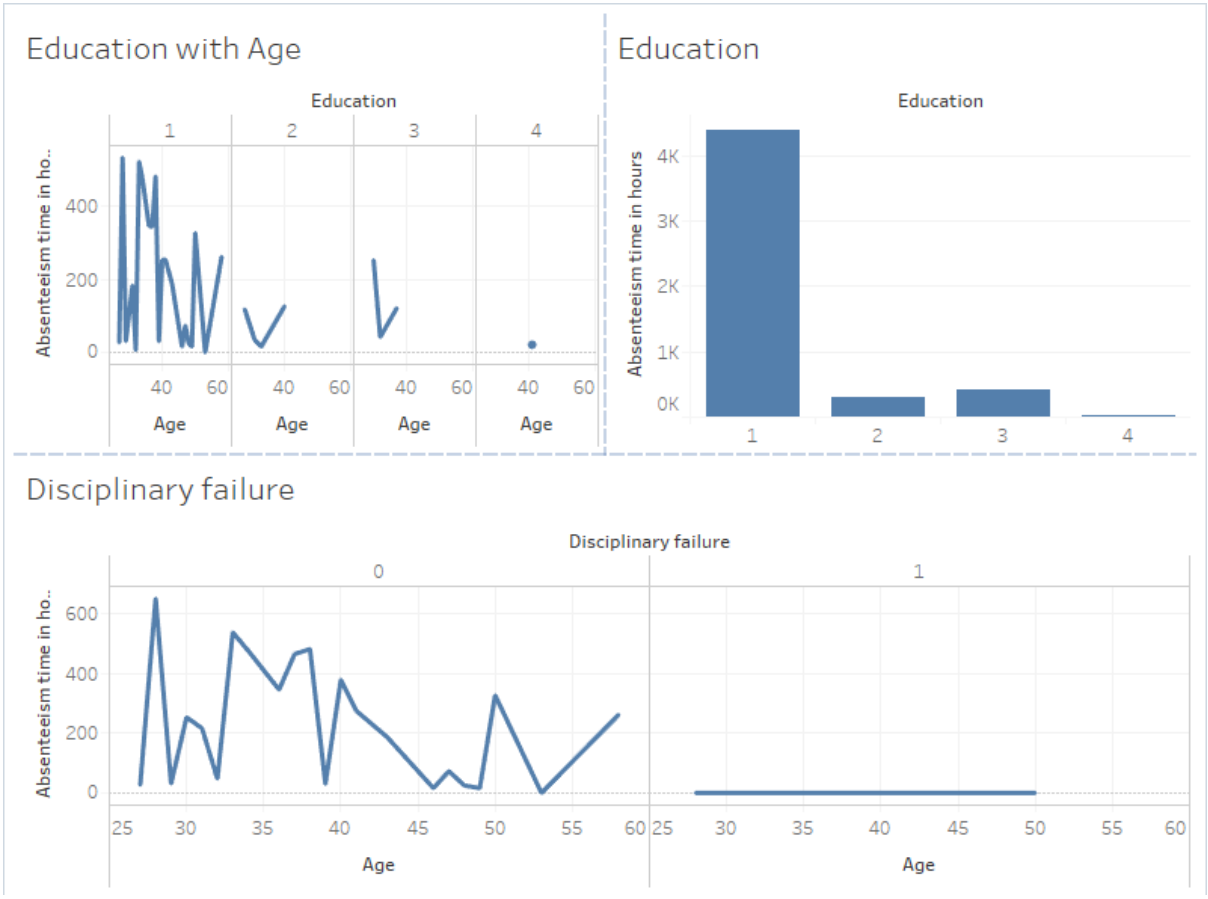
Part II

There are various variables and from above analytics it is confirmed that the following sets of variables have the strongest relationship with each other:

1.) Age and absenteeism
2.) Distance and absenteeism
3.) Reason for absence and absenteeism
4.) Season and absenteeism

We develop dashboards for each of them and explain how they have a strong relationship :

Age and absenteeism



For tracking the relationship, we examine two attributes :

1.) Education with age :
   - Education is inversely related to absenteeism. That means higher educated people tend to miss work less as compared to others.
   - The age group of 20-25 and 40-60 show less absenteeism at work. The youth that is 25 – 25 show more absenteeism.

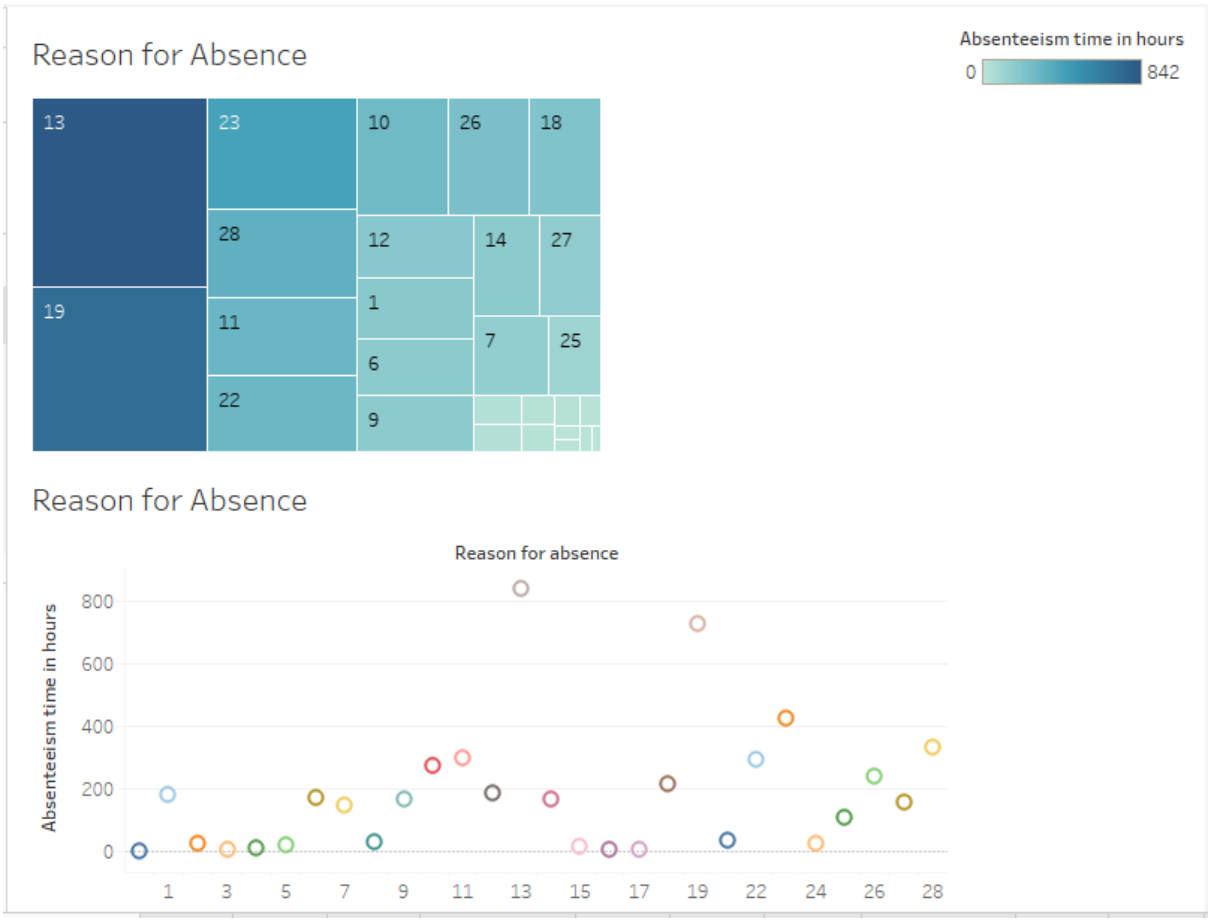
1.) Disciplinary failure with age :
   - Disciplinary failure widely affects the absenteeism .
   - People who are failures tend to miss work more.
   - Again, people with 25-35 do not have disciplinary failure but still show absenteeism at work.

Distance from residence and absenteeism

When it comes to distance, transportation expense is also taken into consideration to see how it affects the absenteeism

1.) An unusual relationship is seen between distance and absenteeism. People living from 25 to 40 kms of work tend to miss work more
2.) Then there is drop in the graph for 40 to 50 kms, and then there is a sudden rise in absenteeism for 50 to 55 kms.
3.) However, to some extent it can be concluded that relationship is linear. When the distance increases absenteeism increases.
4.) An obvious statement can be made that if the distance is more , travel expense will be more.
5.) Hence in the above dashboard we compare the graphs of distance and transportation expense plotted against absenteeism.
6.) We can easily see that the trend in the graph is somewhat similar.
7.) Work load average is also taken into consideration , as considering the travel time of employee the performance at his work will be affected.

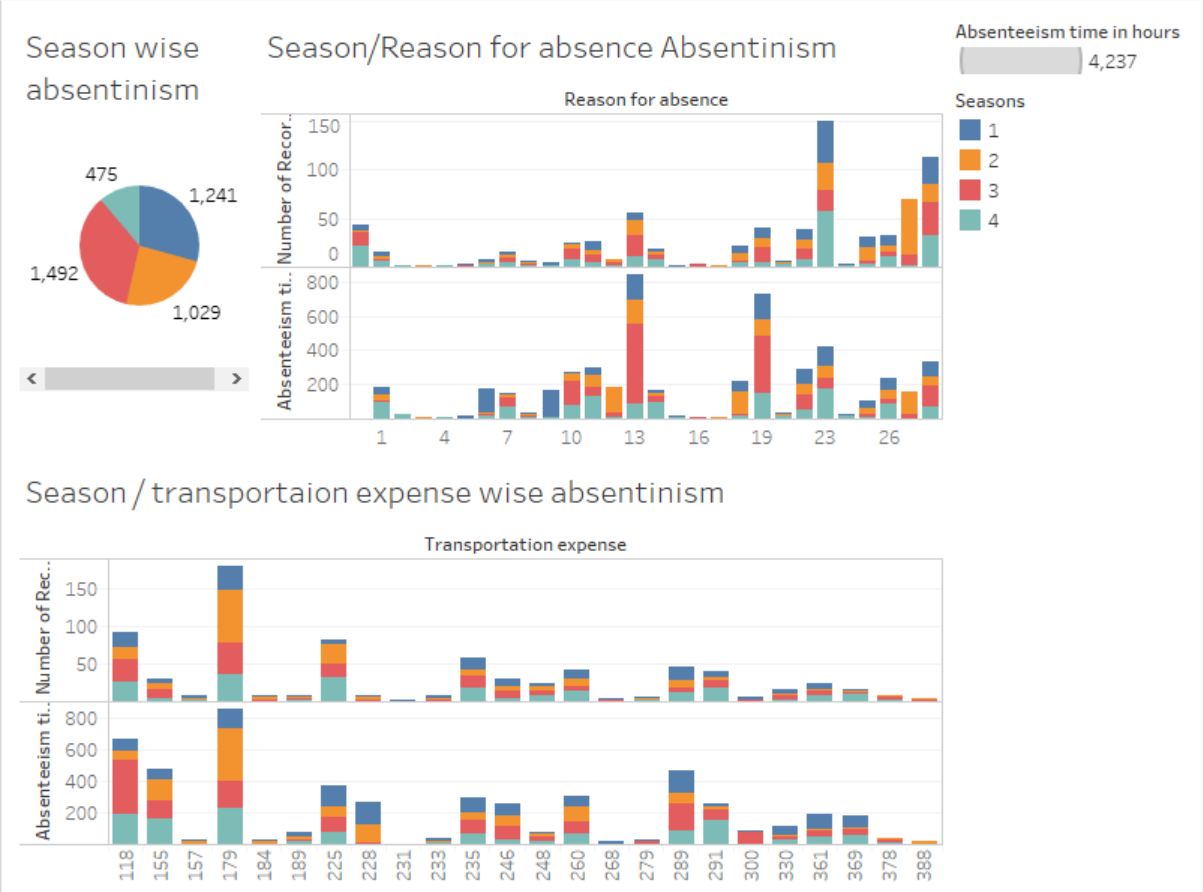
Reason for absence and Absenteeism



We have a set of 28 reasons for absence , out of which 21 are medical conditions . Remaining reason can be unjustified or any other reasons:

1. From the above dashboard it can be concluded that, medical conditions affect more than any other reason.
2. Reason no. 13 (Diseases of the musculoskeletal system and connective tissue) affects the most. This can be caused by accidents while commuting to work or somewhere else.
3. The above is followed by reason no. 19 (Injury, poisoning and certain other consequences of external causes). This can be listed under communicable diseases.
4. Measures can be taken by identifying spreading diseases and ask employees to take precautions accordingly.

Season and Absenteeism



In the above graph, the colours are for seasons , where :
1: Summer
2: Autumn
3: Winter
4: Spring

As we can see from the above graph, most absenteeism is seen in the season of winter. As the weather condition worsens in winter, it can be expected that people prefer staying at home in order to avoid falling sick.

We also, try to see if seasons are related to reasons of absence.=

1.) From the graph it is evident that winter also leads to medical conditions that can cause absence.

Next we also check, if seasons and transportation expense have any relationship between them.

1.) The transportation expense rises in autumn followed by winter.
2.) Very less absenteeism is seen in summer and spring.

Actionable Intelligence:

- Sickness being one of the major cause, employers should pay attention to the physical fitness of the employees
- If any medical condition is recurring, the root cause should be found out and it should be minimized.
- Commute to work should be provided to people who live far off.
- Leaves should be granted after 4-5 days of continuous work.
- Reducing the work pressure by increasing work force can improve the situation and minimize absenteeism.