ANALYSIS OF ABSENTEEISM DATASET

TEAM NAME: CTRL DATA INTELLIGENCE

TEAM MEMBERS

1) MEET PARIKH
2) VIVEK SALUNKHE

COLLEGE NAME: NMIMS Mumbai

Dataset: Analysis of absenteeism in a company

The dataset consists of records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. The dataset initially contains 740 records and 21 features.

Links you must refer for complete and better understanding of the Data Pre-processing and EDA in Python and Tableau

Note: Upon opening the Tableau Visualization link view the story in Full Screen Mode only (Press the full screen button ⛶ present on the bottom right corner of the story page). You can further hover upon any graphs present in the story to find additional information related to it in the form of tooltip.

Content	*Links to access the content*
Data Pre-Processing	https://github.com/VivekSalunkhe14/Cascade/blob/main/Cascade_Preprocessing.ipynb
Exploratory Data Analysis in python	https://github.com/VivekSalunkhe14/Cascade/blob/main/Cascade_EDA.ipynb
Tableau visualization	https://public.tableau.com/profile/vivek1376#!/vizhome/Vivek_Meet_EDA/OverviewofAbsenteeismTimeDataofEmployees?publish=yes

Data Pre-processing:

Prior to performing analysis on the dataset, it is essential to pre-process the data and prepare it in the form to make it suitable for our use.

Steps followed for Data Pre-processing:

1. **Checking the datatypes of features:** Initially the dataset contained all the numerical columns with data type as integer, however we observed that there were few categorical columns present in our dataset. Our first step began with modifying the data types of these columns and assigning it appropriately as category rather than integer. These columns included:

   Reason for absence, Month of absence, Day of the week, Seasons, Education, Disciplinary failure, Social drinker, Social smoker

   Further after completion of this step we also obtained memory optimization in our original dataset where the storage space reduced to **84.1kB** from **121.5kB**.

2. **Checking for NA Values:** We then performed a check for null values in the dataset. The pre-processing step confirmed that data has all valid inputs and have no such NA or Nan values throughout our data.

3. **Handling invalid data in Reason for Absence feature:** Occasionally the inserted information does not explain the exact information of features. Similar issue was encountered in the feature named **“Reason for absence”** which can take any integer value in between 1 to 28 and each individual value is assigned to specified reason. However, we observed some inputs where 0 was mentioned as the reason for absence which clearly states that information is either misinterpreted or it is a human error. For proper analysis, all the data points 0 as the Reason for absence was replaced with value 26. This was performed after identifying that the Reason 26th indicated **Unjustified Absence** which seems suitable to be replaced.

4. **Mapping value of character attributes to its pre-defined categories:**

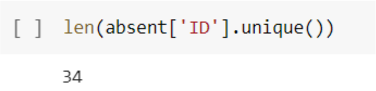
   Since our dataset had numeric values in the categorical features it would be difficult to interpret those columns while performing Exploratory Data Analysis. So, with the help of data description, the categorical features namely **Reason for absence, Month of absence, Day of the week, Season, Education, Disciplinary failure, Social drinker, Social smoker** were mapped to its corresponding pre-defined categorical labels. The newly created attributes were as follows:

   Reason_Justified, Month_Name, Day_Name, Season_Name, Education_Details, Disciplinary_Status, Drinking_Status, Smoking_Status

5. **Inconsistency in Month column:** Feature **Month of absence** can take any value between 1 to 12 as it suggests month number for a particular year, but 0 value clearly dictates error in datapoints which needs to be dropped as it would be incorrect to replace this value.

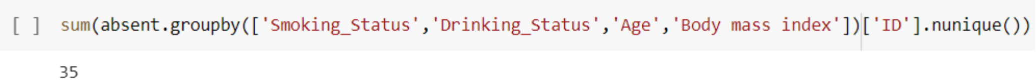
6. **Inconsistency in ID column:**

a. Initially we observed that there were 34 Unique ID's in the dataset



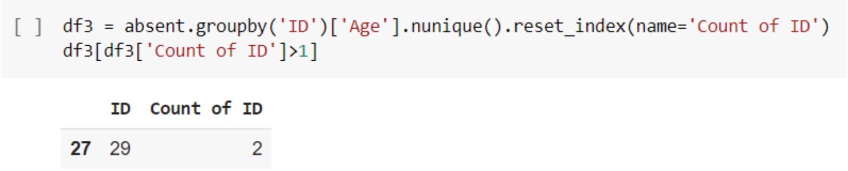
```
[ ]  len(absent['ID'].unique())

     34
```

b. However, after grouping the datasets based on some essential features, we observed that there was discrepancy in count of ID's and one of the ID was duplicated



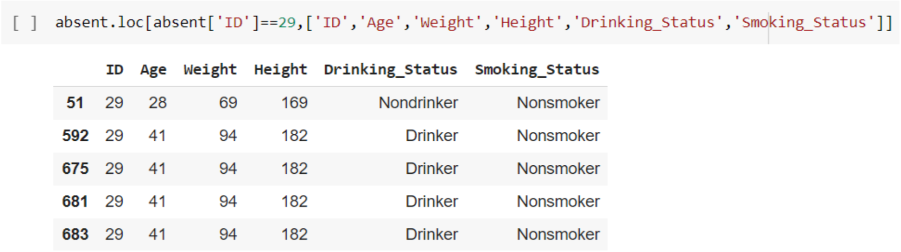
```
[ ]  sum(absent.groupby(['Smoking_Status','Drinking_Status','Age','Body mass index'])['ID'].nunique())

     35
```

c. In order to identify the extra duplicated ID record, we grouped the dataset by ID and counted the unique values of Age. Once this was done, we extracted the recorded containing count of ID greater than 1. We observed that there was an ID 29 present with count of 2 which clearly indicates that one of the records with ID 29 is incorrect since all the similar ID will have same Age as they indicate the same individual.



```
[ ]  df3 = absent.groupby('ID')['Age'].nunique().reset_index(name='Count of ID')
     df3[df3['Count of ID']>1]

         ID  Count of ID
     27  29            2
```

d. After extracting the data of 29th ID we found out that there was one record with Age 28 which seems incorrect as all the other datapoints were having Age 41 for 29th ID. We then drop that particular inconsistent record from our dataset

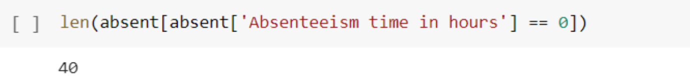


```
[ ]  absent.loc[absent['ID']==29,['ID','Age','Weight','Height','Drinking_Status','Smoking_Status']]

          ID  Age  Weight  Height  Drinking_Status  Smoking_Status
     51   29   28      69     169       Nondrinker       Nonsmoker
     592  29   41      94     182          Drinker       Nonsmoker
     675  29   41      94     182          Drinker       Nonsmoker
     681  29   41      94     182          Drinker       Nonsmoker
     683  29   41      94     182          Drinker       Nonsmoker
```

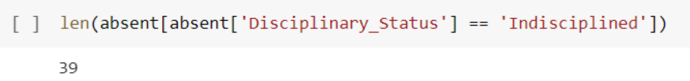
7. Inconsistency in Absenteeism time in hours column:

a. Dataset aims to provide information for absenteeism but 0 value in **"Absenteeism time in hours"** indicates discrepancy in data and we need to specify certain value for such records as per constraints to remove contradiction.

b. There exist 40 rows for 0 absenteeism time but only 39 rows for Indisciplinary status. One row must be valid which states that ID is disciplined and having valid reason. Let us identify that particular valid row.

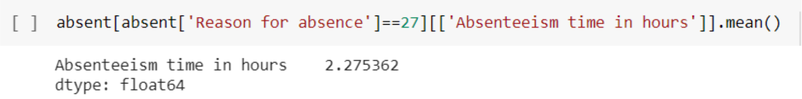


```
[ ]  len(absent[absent['Absenteeism time in hours'] == 0])

     40
```



```
[ ]  len(absent[absent['Disciplinary_Status'] == 'Indisciplined'])

     39
```

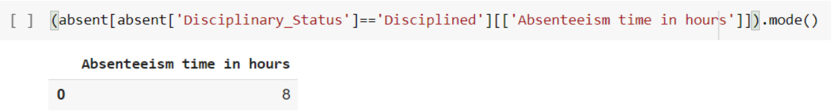
c. We found out that one valid record had ID 27 so for this particular one datapoint the Absenteeism time was replaced with average value of absent time for the reason stated by that corresponding ID 27 i.e., due to Physiotherapy.



```
[ ]  absent[absent['Reason for absence']==27][['Absenteeism time in hours']].mean()

     Absenteeism time in hours    2.275362
     dtype: float64
```

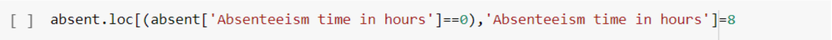
d. After replacing that particular ID, we are left with 39 datapoints whose absenteeism time was left as 0

e. It is assumed that absenteeism time for those 39 data points will be mode (maximum occurring number) value of all Disciplined data absenteeism time. So, 8 hours is assigned for proper EDA analysis.



```
[ ]  (absent[absent['Disciplinary_Status']=='Disciplined'][['Absenteeism time in hours']]).mode()
```

	Absenteeism time in hours
0	8



```
[ ]  absent.loc[(absent['Absenteeism time in hours']==0),'Absenteeism time in hours']=8
```

f. After removing all the inconsistencies, we are left with 0 records with absenteeism time having 0 value which indicates dataset now contains valid data

8. **Extracting useful information by using calculation of features:** In order to perform efficient analysis of the data, we created certain new features based on the existing features. The new features computed were as follows:

a. $\text{Travel Expense per km} = \frac{Transportation\ Expense}{Distance\ from\ Residence\ to\ Work}$

b. $Travel\ Cost\ Category$:
$\text{Travel Expense per km} < 7 \rightarrow Cheap$
$7 \ \leq \text{Travel Expense per km} \ \leq 15 \rightarrow Affordable$
$Travel\ Expense\ per\ km > 15 \ \rightarrow Expensive$

c. $Joining\ Age = Age - Service\ Time$

d. Age_Group:
$\text{Age} < 35 \rightarrow Young\ Employee$
$35 \leq \text{Age} \leq 45 \rightarrow Mid - Age\ Employee$
$\text{Age} > 45 \rightarrow Old\ Employee$

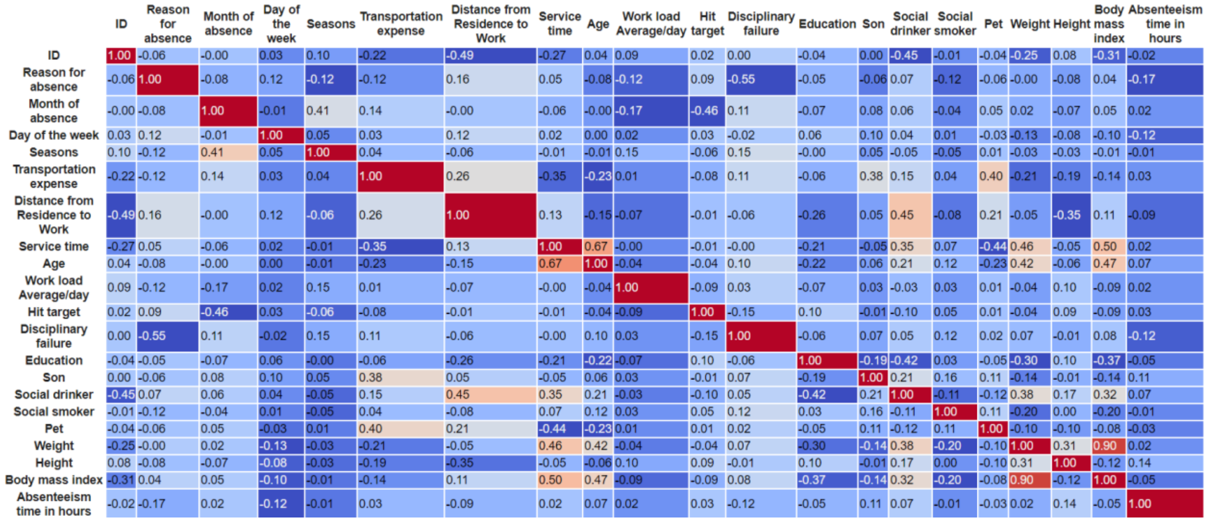
e. BMI_status:
$\text{Body mass index} < 18.5 \rightarrow Underweight$
$18.5 \leq \text{Body mass index} < 24.9 \rightarrow Normal$
$24.9 \leq \text{Body mass index} \leq 29.9 \rightarrow Overweight$
$\text{Body mass index} > 29.9 \rightarrow Obese$

Exploratory Data Analysis: Once the pre-processing step was performed, we imported both the pre-processed and the original dataset for our Exploratory Data Analysis.

Data Visualizations:

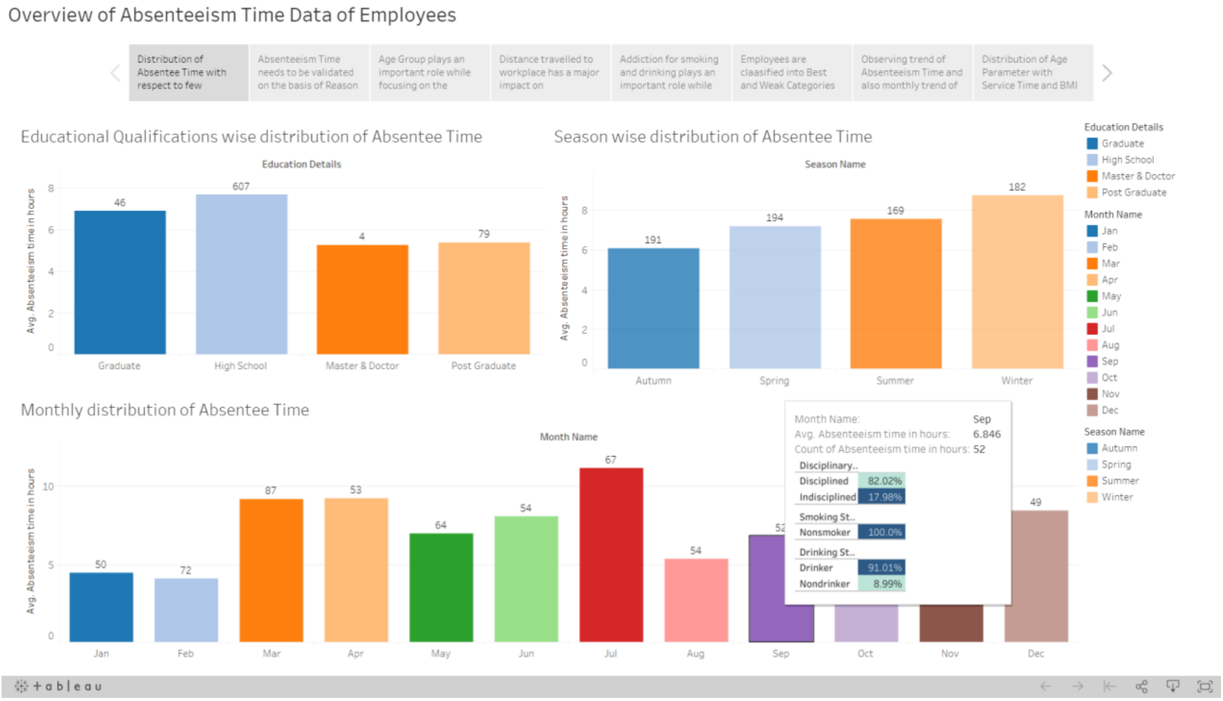
1. Correlation Plot:

Initially we plot the correlation matrix for our original dataset. We observed that majority of the correlation values are near to 0 which states that there are no highly dependent features except some of the features having value near to ±0.45 which indicates slightly correlated attributes.



	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
ID	1.00	-0.06	-0.00	0.03	0.10	-0.22	-0.49	-0.27	0.04	0.09	0.02	0.00	-0.04	0.00	-0.45	-0.01	-0.04	-0.25	0.08	-0.31	-0.02
Reason for absence	-0.06	1.00	-0.08	0.12	-0.12	-0.12	0.16	0.05	-0.08	-0.12	0.09	-0.55	-0.05	-0.06	0.07	-0.12	-0.06	-0.00	-0.08	0.04	-0.17
Month of absence	-0.00	-0.08	1.00	-0.01	0.41	0.14	-0.00	-0.06	-0.00	-0.17	-0.46	0.11	-0.07	0.08	0.06	-0.04	0.05	0.02	-0.07	0.05	0.02
Day of the week	0.03	0.12	-0.01	1.00	0.05	0.03	0.12	0.02	0.00	0.02	0.03	-0.02	0.06	0.10	0.04	0.01	-0.03	-0.13	-0.08	-0.10	-0.12
Seasons	0.10	-0.12	0.41	0.05	1.00	0.04	-0.06	-0.01	-0.01	0.15	-0.06	0.15	-0.00	0.05	-0.05	-0.05	0.01	-0.03	-0.03	-0.01	-0.01
Transportation expense	-0.22	-0.12	0.14	0.03	0.04	1.00	0.26	-0.35	-0.23	0.01	-0.08	0.11	-0.06	0.38	0.15	0.04	0.40	-0.21	-0.19	-0.14	0.03
Distance from Residence to Work	-0.49	0.16	-0.00	0.12	-0.06	0.26	1.00	0.13	-0.15	-0.07	-0.01	-0.06	-0.26	0.05	0.45	-0.08	0.21	-0.05	-0.35	0.11	-0.09
Service time	-0.27	0.05	-0.06	0.02	-0.01	-0.35	0.13	1.00	0.67	-0.00	-0.01	-0.00	-0.21	-0.05	0.35	0.07	-0.44	0.46	-0.05	0.50	0.02
Age	0.04	-0.08	-0.00	0.00	-0.01	-0.23	-0.15	0.67	1.00	-0.04	-0.04	0.10	-0.22	0.06	0.21	0.12	-0.23	0.42	-0.06	0.47	0.07
Work load Average/day	0.09	-0.12	-0.17	0.02	0.15	0.01	-0.07	-0.00	-0.04	1.00	-0.09	0.03	-0.07	0.03	-0.03	0.03	0.01	-0.04	0.10	-0.09	0.02
Hit target	0.02	0.09	-0.46	0.03	-0.06	-0.08	-0.01	-0.01	-0.04	-0.09	1.00	-0.15	0.10	-0.01	-0.10	0.05	0.01	-0.04	0.09	-0.09	0.03
Disciplinary failure	0.00	-0.55	0.11	-0.02	0.15	0.11	-0.06	-0.00	0.10	0.03	-0.15	1.00	-0.06	0.07	0.05	0.12	0.02	0.07	-0.01	0.08	-0.12
Education	-0.04	-0.05	-0.07	0.06	-0.00	-0.06	-0.26	-0.21	-0.22	-0.07	0.10	-0.06	1.00	-0.19	-0.42	0.03	-0.05	-0.30	0.10	-0.37	-0.05
Son	0.00	-0.06	0.08	0.10	0.05	0.38	0.05	-0.05	0.06	0.03	-0.01	0.07	-0.19	1.00	0.21	0.16	0.11	-0.14	-0.01	-0.14	0.11
Social drinker	-0.45	0.07	0.06	0.04	-0.05	0.15	0.45	0.35	0.21	-0.03	-0.10	0.05	-0.42	0.21	1.00	-0.11	-0.12	0.38	0.17	0.32	0.07
Social smoker	-0.01	-0.12	-0.04	0.01	-0.05	0.04	-0.08	0.07	0.12	0.03	0.05	0.12	0.03	0.16	-0.11	1.00	0.11	-0.20	0.00	-0.20	-0.01
Pet	-0.04	-0.06	0.05	-0.03	0.01	0.40	0.21	-0.44	-0.23	0.01	0.01	0.02	-0.05	0.11	-0.12	0.11	1.00	-0.10	-0.10	-0.08	-0.03
Weight	-0.25	-0.00	0.02	-0.13	-0.03	-0.21	-0.05	0.46	0.42	-0.04	-0.04	0.07	-0.30	-0.14	0.38	-0.20	-0.10	1.00	0.31	0.90	0.02
Height	0.08	-0.08	-0.07	-0.08	-0.03	-0.19	-0.35	-0.05	-0.06	0.10	0.09	-0.01	0.10	-0.01	0.17	0.00	-0.10	0.31	1.00	-0.12	0.14
Body mass index	-0.31	0.04	0.05	-0.10	-0.01	-0.14	0.11	0.50	0.47	-0.09	-0.09	0.08	-0.37	-0.14	0.32	-0.20	-0.08	0.90	-0.12	1.00	-0.05
Absenteeism time in hours	-0.02	-0.17	0.02	-0.12	-0.01	0.03	-0.09	0.02	0.07	0.02	0.03	-0.12	-0.05	0.11	0.07	-0.01	-0.03	0.02	0.14	-0.05	1.00

2. Distribution of Absentee Time with respect to few important parameters like Educational Qualification, Season and Month of the Year



Overview of Absenteeism Time Data of Employees
Distribution of Absentee Time with respect to few
Absenteeism Time needs to be validated on the basis of Reason
Age Group plays an important role while focusing on the
Distance travelled to workplace has a major impact on
Addiction for smoking and drinking plays an important role while
Employees are claasified into Best and Weak Categories
Observing trend of Absenteeism Time and also monthly trend of
Distribution of Age Parameter with Service Time and BMI
Educational Qualifications wise distribution of Absentee Time
Education Details
Avg. Absenteeism time in hours
0
2
4
6
8
46
607
4
79
Graduate
High School
Master & Doctor
Post Graduate
Season wise distribution of Absentee Time
Season Name
191
194
169
182
Autumn
Spring
Summer
Winter
Education Details
Graduate
High School
Master & Doctor
Post Graduate
Month Name
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Season Name
Autumn
Spring
Summer
Winter
Monthly distribution of Absentee Time
Month Name
0
5
10
50
72
87
53
64
54
67
54
52
49
+tableau

Month Name: Sep
Avg. Absenteeism time in hours: 6.846
Count of Absenteeism time in hours: 52

Disciplinary..	
Disciplined	82.02%
Indisciplined	17.98%

Smoking St..	
Nonsmoker	100.0%

Drinking St..	
Drinker	91.01%
Nondrinker	8.99%

The graph indicates the bar chart between the Absenteeism time in hours with the Important Parameters. The value present on the tip of each bar indicates the total count of the individuals. Upon hovering any particular bar, we get to see information in the form of tooltip which specifies the Average Absenteeism Time in hour as well as it gives us the understanding of the percentage wise distribution of attributes like Disciplinary Status, Smoking Status and Drinking Status.

- Educational Qualification wise Absenteeism Time

Insight: Majority of the High School Graduates were observed to be absent for most of the time.

Proposal: This can be tackled by conducting counselling sessions or workshops in order to make them realise the importance of the job.

- Season wise Absenteeism Time

Insight: Majority of the Absenteeism time was recorded in the Winter and Spring Season.

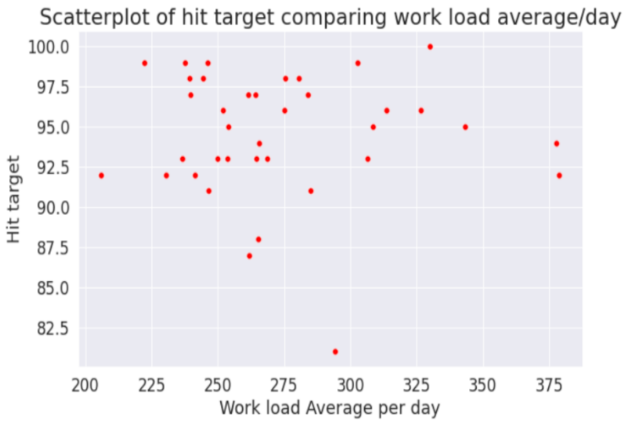
Proposal: In order to overcome the issue company can initiate certain awareness to ensure healthy environment at the workplace.

- Monthly distribution of Absenteeism Time

Insight: We observe that during the months of June and July i.e., during Winter season in Brazil the average of Absenteeism Time is greater as compared to other months. Also, the lowest value is recorded in the months of January and February.

Proposal: Execution of seasonal measures would aid in lowering the monthly absenteeism time

3. Scatterplot of Hit Target and Workload Average Per Day:



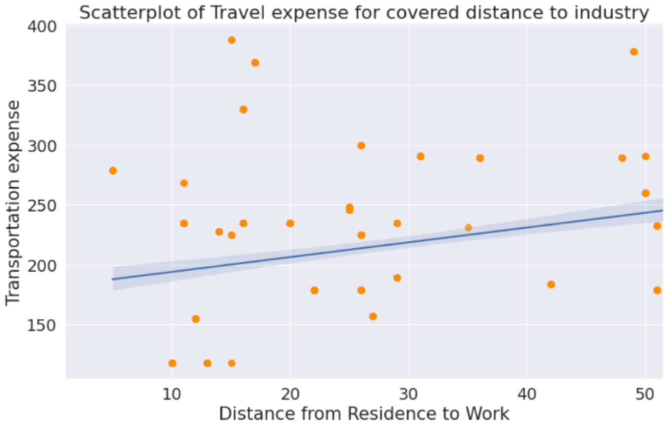
Insight: The graph clearly indicates that as the Workload goes on increasing, we observed a significant fall in the target hit. Also, it can be seen that when the workload was below 275 then majority of the employees hit the target.

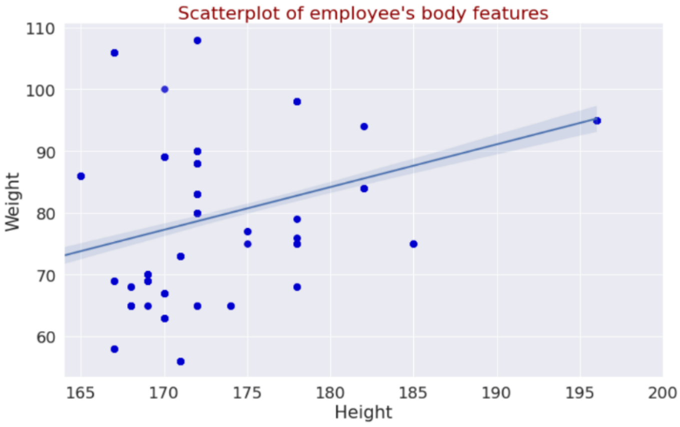
Proposal: The company should focus on evenly distribution of workload within the employees so as to ensure improvement in Hit Target. Also, in order to motivate the employees' company can provide perks for hitting the maximum target either on a monthly or quarterly basis.

4. Scatterplots of certain attributes with Linear Regression line of best fit:

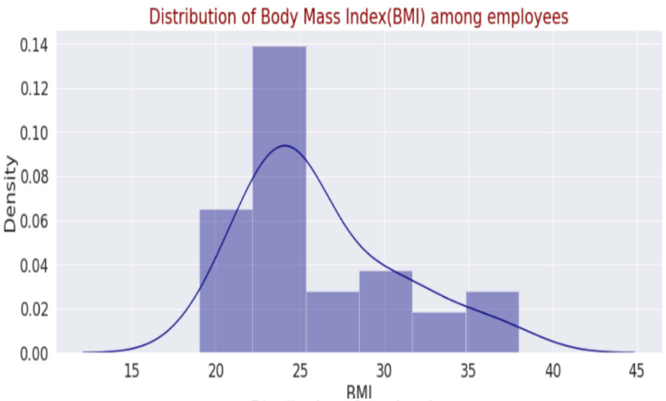
- Distribution of Transportation Expense to Distance from Residence to Work

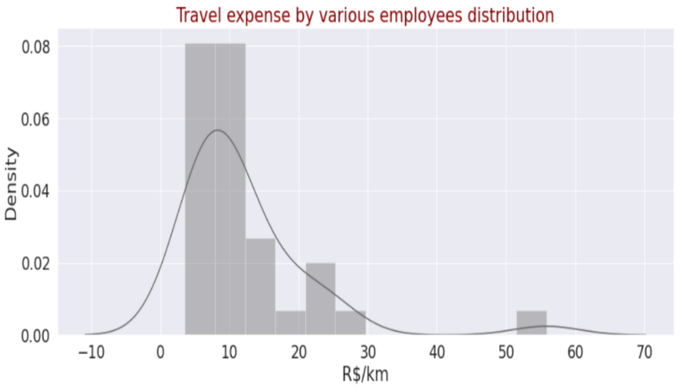
Insight: The graph depicts that there exists linear relationship between the transportation expense and distance from residence to work. Further the positive slope of Linear regression line indicates that as the distance from residence to work increases then the corresponding transportation expense also increases

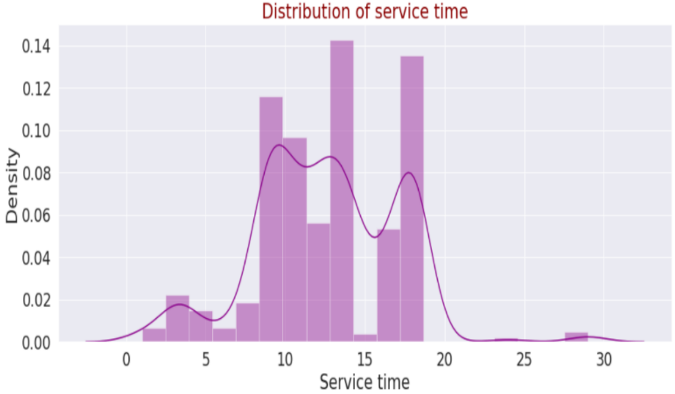


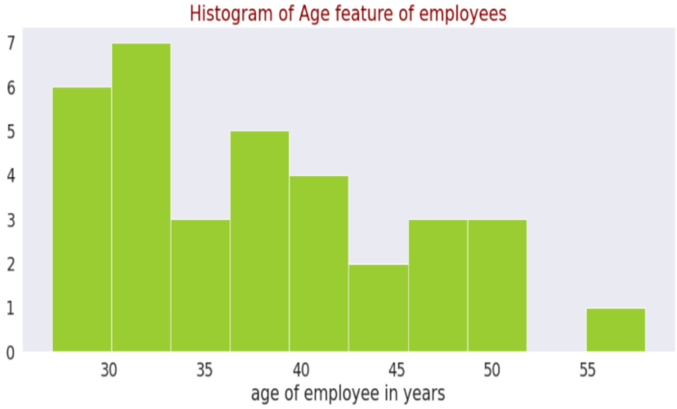


5. Density curves with Histogram









As there were several ID's repeating in our dataset so plotting the histogram with repetitive ID will not be feasible. In order to overcome this, we grouped the data on the basis of ID and plotted the histogram for the several attributes.

- Distribution of BMI among Employees

Insight: It can be clearly visualized from the distribution that there are no employees with BMI less than 19 which indicates that none of the employee fall in the Underweight Category. Also, majority of them fall in the Normal category between 19 to 25 which signifies healthy lifestyle.

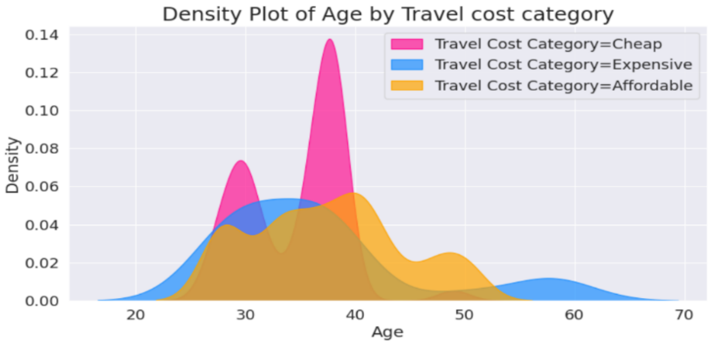
- Distribution of Total Expense per km

Insight: As observed from the distribution majority of the employees spend approximately less than 10R$ per km to travel to the workplace.

- Distribution of Service Time

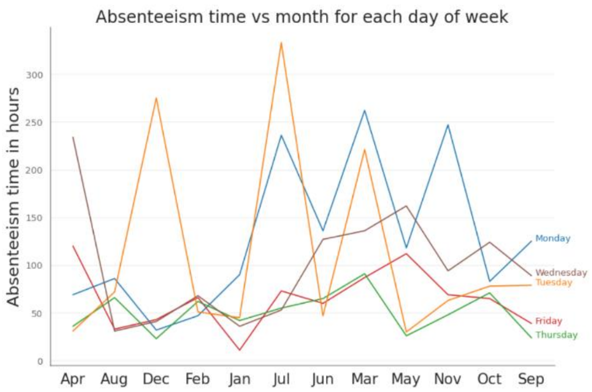
Insight: As seen from the probability density curve that majority of the employees are highly experienced as they lie in the period of 10 to 20 years

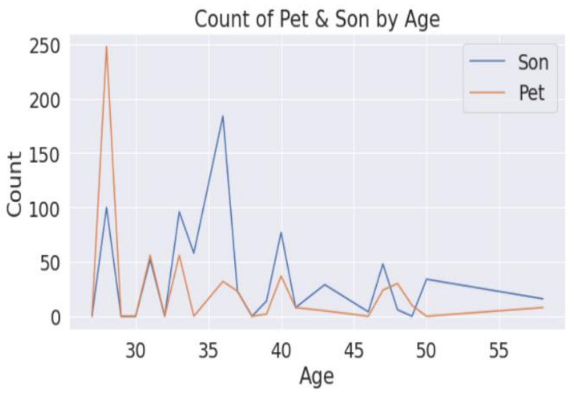
6. Density Plot of Age by Travel Cost Category



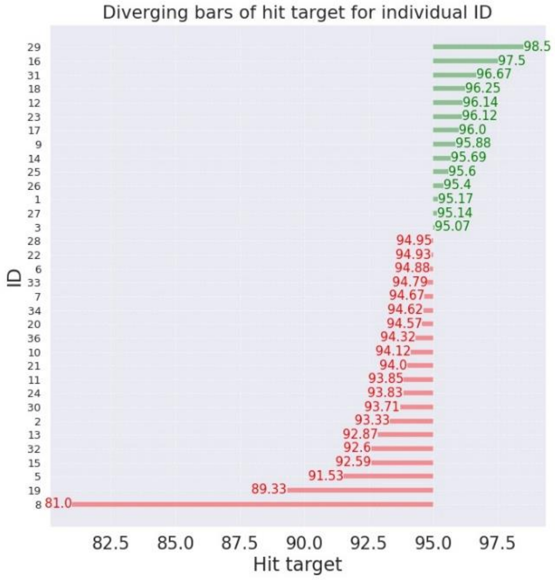
Insight: The graph clearly indicates that majority of the Young and Mid-Age employees belong to Cheap and Affordable Transportation Category and very few prefer Expensive travel. Also, Old Employees focus on comfort and ease while transportation thus they belong to Affordable or Expensive categories.

7. **Seasonal Plot and Trend Line Chart:** Trends between monthly and daily distribution of Absenteeism time in hours and also count of Pet and Son possessed in terms of Age





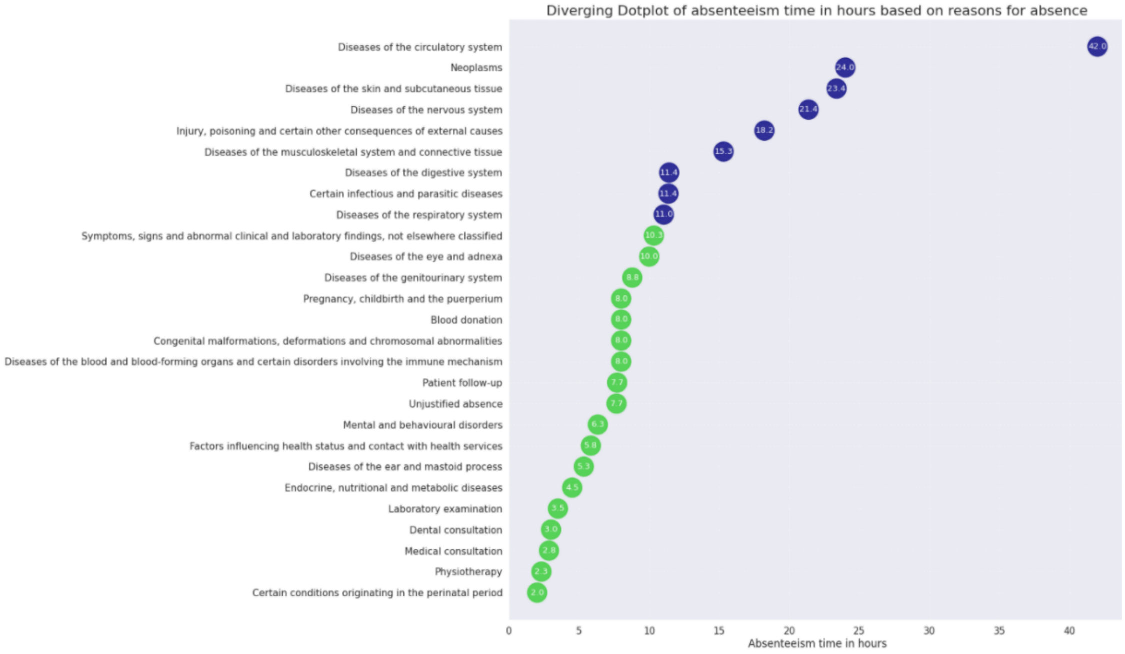
8. Diverging Text Chart



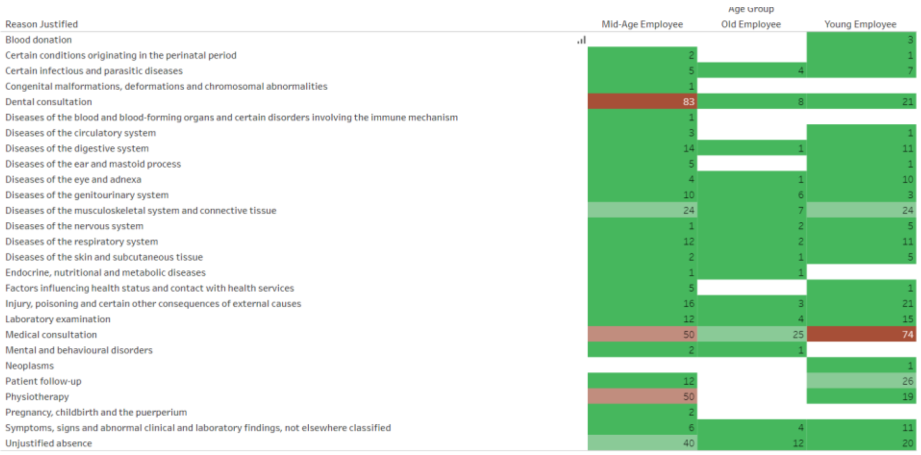
Insight: The graph shows the individuals performing better and worse in terms of completed task. The Green colored ID’s indicate those individuals that are performing good whereas Red colored ID’s gives us an idea about the individuals lacking in terms of work assigned to them.

Proposal: Company need to analyze the causes for underperforming individuals. These causes can be lack in skillset, negative corporate culture, or personal issue. Overcoming these causes would definitely help in improvising the growth of the company.

9. Diverging Dot Plot:



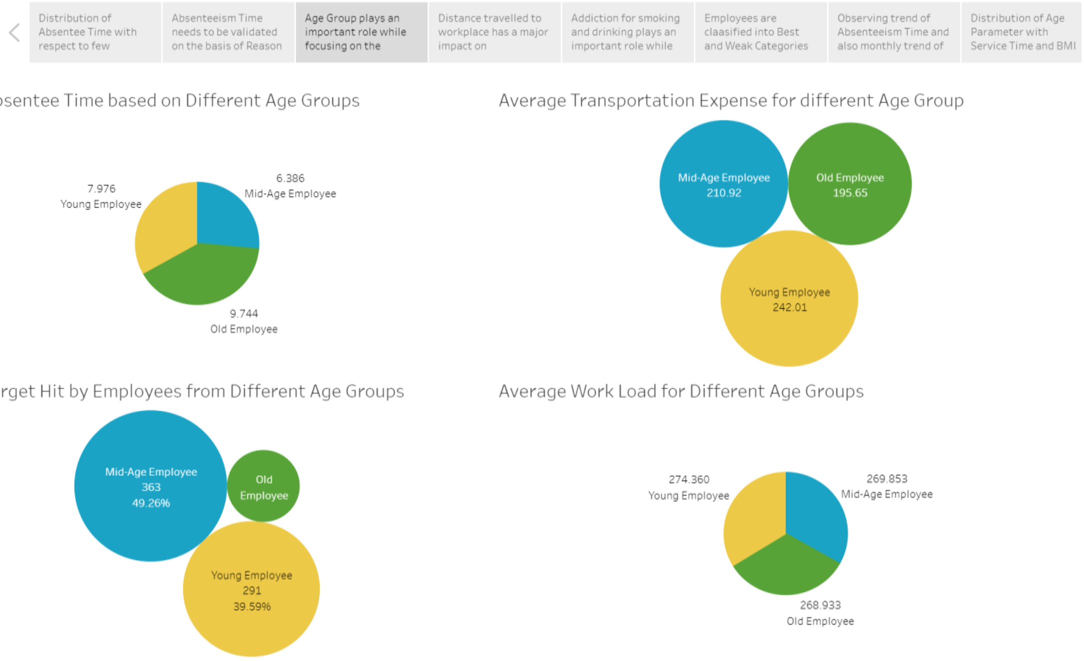
Insight: This graph gives an idea about the Reason of Absence in comparison to Average Absenteeism Time in hours.
Proposal: This will help to identify the critical reason having maximum average value in Absenteeism Time for which the company can initiate some counter measure to overcome them



Reason Justified	Age Group Mid-Age Employee	Old Employee	Young Employee
Blood donation			3
Certain conditions originating in the perinatal period	2		1
Certain infectious and parasitic diseases	5	4	7
Congenital malformations, deformations and chromosomal abnormalities	1		
Dental consultation	83	8	21
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	1		
Diseases of the circulatory system	3		1
Diseases of the digestive system	14	1	11
Diseases of the ear and mastoid process	5		1
Diseases of the eye and adnexa	4	1	10
Diseases of the genitourinary system	10	6	3
Diseases of the musculoskeletal system and connective tissue	24	7	24
Diseases of the nervous system	1	2	5
Diseases of the respiratory system	12	2	11
Diseases of the skin and subcutaneous tissue	2	1	5
Endocrine, nutritional and metabolic diseases	1	1	
Factors influencing health status and contact with health services	5		1
Injury, poisoning and certain other consequences of external causes	16	3	21
Laboratory examination	12	4	15
Medical consultation	50	25	74
Mental and behavioural disorders	2	1	
Neoplasms			1
Patient follow-up	12		26
Physiotherapy	50		19
Pregnancy, childbirth and the puerperium	2		
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	6	4	11
Unjustified absence	40	12	20

Insight: This graph gives an idea about the critical reason specified for absence by the individuals in different Age Groups. Few of the critical reasons include medical consultation, dental consultation, physiotherapy, etc.

10. Packed Bubble Chart and Pie Chart for several features with respect to Age Group


Distribution of Absentee Time with respect to few
Absenteeism Time needs to be validated on the basis of Reason
Age Group plays an important role while focusing on the
Distance travelled to workplace has a major impact on
Addiction for smoking and drinking plays an important role while
Employees are classified into Best and Weak Categories
Observing trend of Absenteeism Time and also monthly trend of
Distribution of Age Parameter with Service Time and BMI
Average Absentee Time based on Different Age Groups
7.976
Young Employee
6.386
Mid-Age Employee
9.744
Old Employee
Average Transportation Expense for different Age Group
Mid-Age Employee
210.92
Old Employee
195.65
Young Employee
242.01
Average Target Hit by Employees from Different Age Groups
Mid-Age Employee
363
49.26%
Old
Employee
Young Employee
291
39.59%
Average Work Load for Different Age Groups
274.360
Young Employee
269.853
Mid-Age Employee
268.933
Old Employee

Age Group plays an important role while focusing on the behaviour of an individual. This graph depicts the distribution of Age Group with major parameters like Absenteeism Time, Transportation Expense, Hit Target, and Workload

- Average Absenteeism Time based on Age Group

Insight: It can be observed that Old Employee tend to remain absent higher than that of Young and Mid-Age Employee. This graph shows the obvious result since the Old Employees are more susceptible to disease and they tend to possess less immunity.

- Average Transportation Expense based on Age Group

Insight: It can be viewed that Young Employee tend to spend more for travelling. Further the Mid-Age and Old Employee spend comparatively less.

- Average Target Hit based on Age Group

Insight: It can be observed that the count of Young and Mid Age Employee is greater than that of Old Employee. However, the average target hit by all the age groups is more or less the same.

- Average Workload based on Age Group

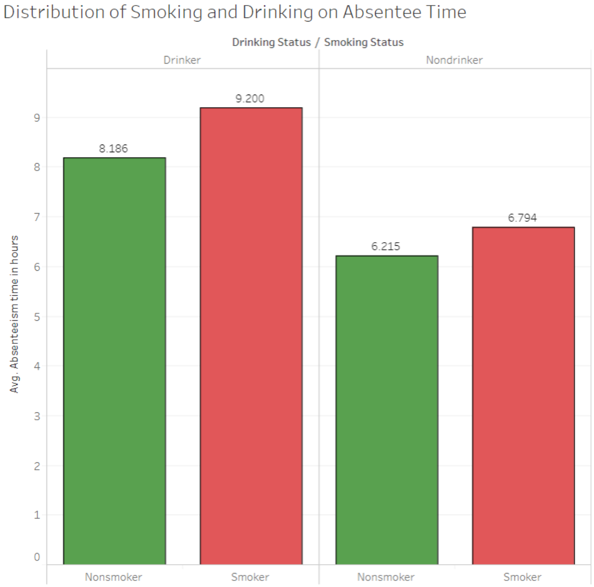
Insight: This graph indicates that the workload between all the Age categories is evenly distributed.

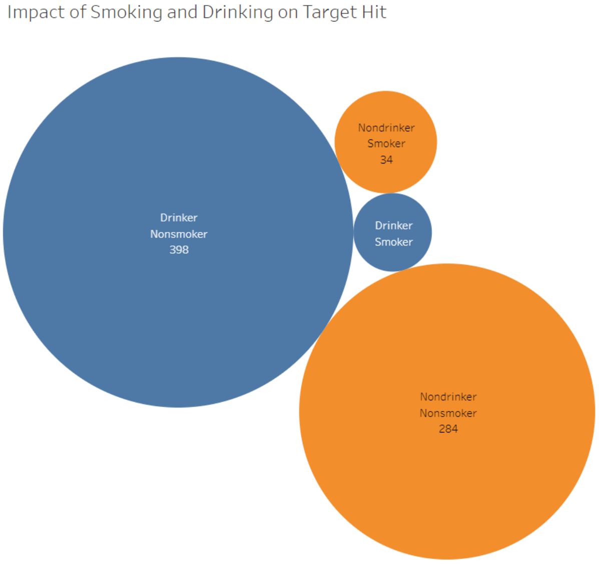
11. Bar Chart and Packed Bubble Chart indicating Negative Impacts of Addiction

Addiction for smoking and drinking plays an important role while analysing the Work performed by an Individual as well as the absenteeism time in hours

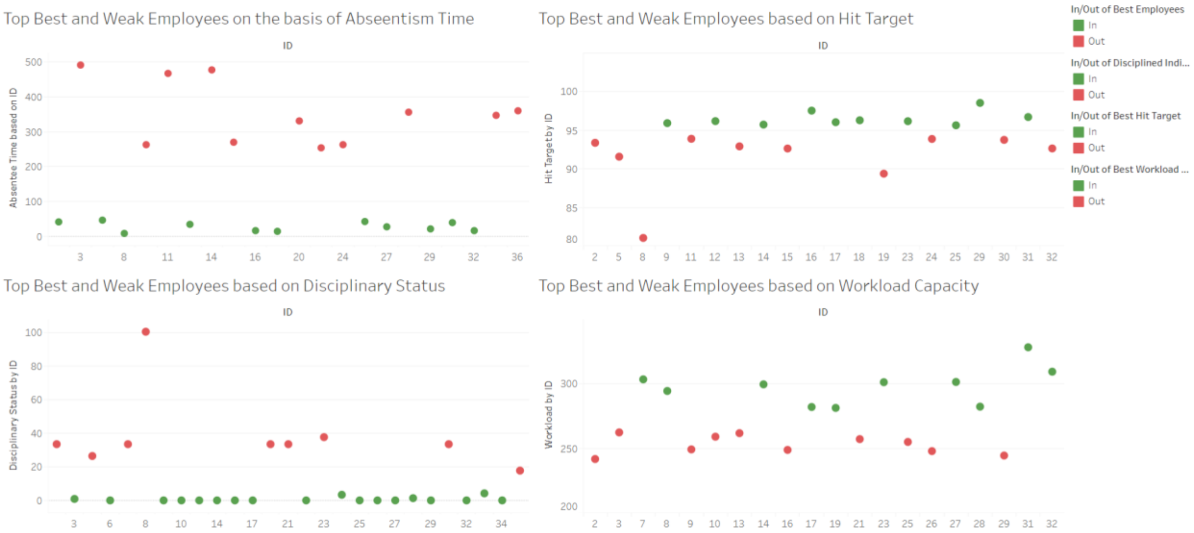
- Distribution of Smoking and Drinking on Absenteeism Time and Hit Target

Insights: It can be observed that the individuals addicted to smoking and drinking have highly negative impact on their work life while the employee having no addiction showed satisfactory performance.





12. Dot plot for Best and Weak Employees based on major attributes



Insight: Employees are classified into Best and Weak Categories based on the different parameters of interests. This helps analysing the weak employees and providing necessary trainings in order to enhance productivity of individual and in turn improve growth of the industry. The Red points in the graph indicates the Weak Employees which needs to be focused upon whereas Green points indicates the best employees which can further be appreciated or rewarded to give consistent performance.

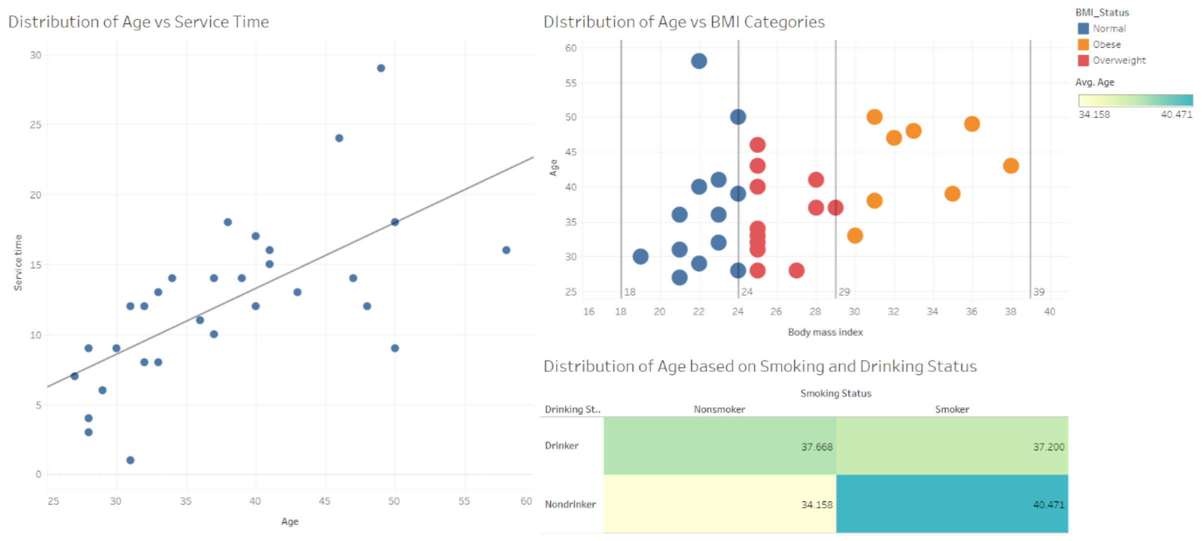
13. Distribution of Age Parameter with Service Time and BMI Categories of different individuals

- Distribution of Age based on Service Time

Insight: The scatter plot indicates that Employees tends to stick with the industry for longer duration which further shows the sign that the industry is treating their employees in a great manner.

- Distribution of Age based on BMI Categories

Insight: The graph indicates that majority of the Young Age Employee belong to Normal and Overweight Category whereas few Mid-Age and Old Employee belong to Obese Category



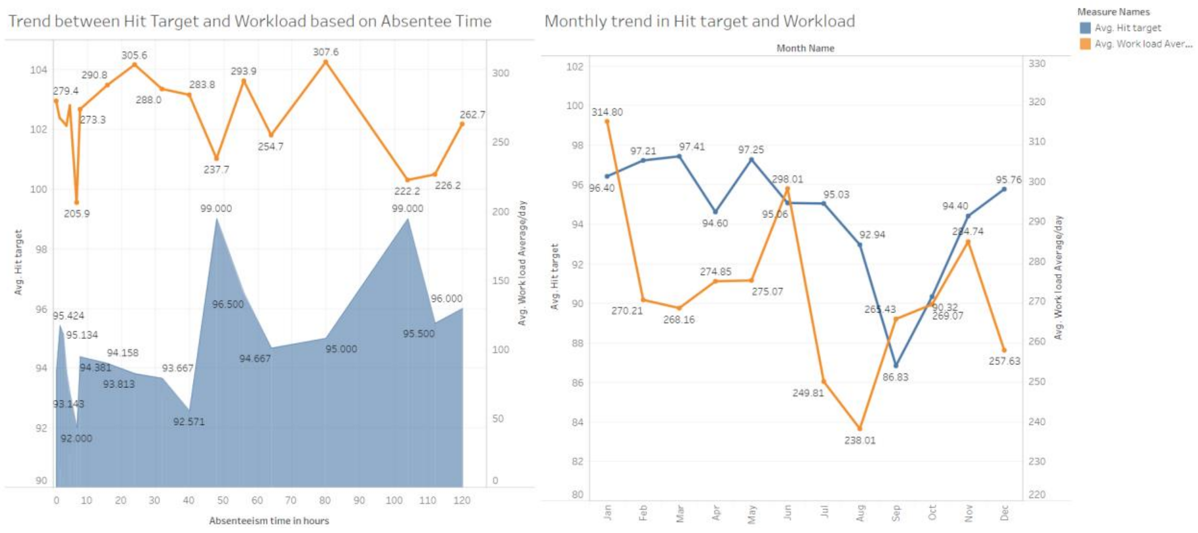
14. Area Chart and Trend Line Chart between several attributes

- Trend between Hit Target, Workload and Absenteeism Time

Insight: Observing trend of Absenteeism Time and also monthly trend of measuring factors like Hit Target and Workload Capacity

- Monthly Trend in Hit Target and Workload

Insight: This graph depicts the inverse relationship between the Workload and Target Hit.



THANK YOU!