Cascade Cup Round 3

Absenteeism Data Analysis

By Third Degree Burn

Aryan Pandey, Nihal John George (IIT Madras)

INSIGHTS

1. Incorrect Data Entry and Some columns are constant in time

Most columns like Age, Service time, distance to work, education etc remain constant for all instances of a particular ID.

Since this data spans over 3 years, we can assume that the data for these constant columns was fed in from an employee database at the time this data was provided.

That is, when an absenteeism instance takes place, only Reason for absence, ID, Month of absence, Day of the week, Seasons, Work load, Disciplinary failure and Absenteeism time in hours are entered into the record.

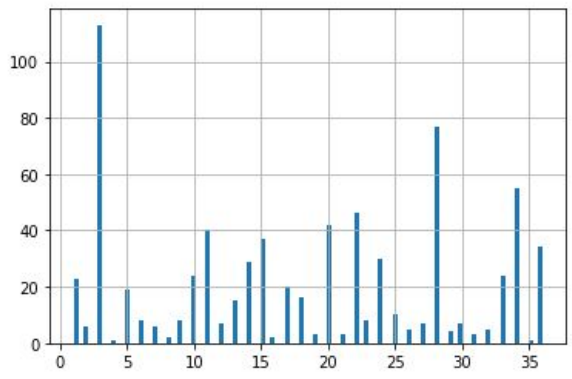
These are the only columns whose data varies from row to row for the same ID.

The other data may have been filled later from a separate employee database with personal details, which causes all their values to take the latest value in that database. This value may have been taken at time of joining and probably never updated too.

Using this insight, we make the following observations

a. There are 36 employees in this dataset (due to 36 unique IDs)

```
In [23]: plt.figure()
         data['ID'].hist(bins = 100)
         plt.show()
```



b. Row 51 with ID 29 is anomalous

Row 51 shown below has an anomalous record of absenteeism for ID 29. This is the only ID and the only row to have a difference in service time, age, distance to work, education, weight, height etc.

```
In [18]: data[data['ID']==29].head()
```

Out[18]:

	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	...	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	W
51	29	0	9	2	4	225	26	9	28	241.476	...	1	1	1	0	0	2	
592	29	28	2	6	2	225	15	15	41	264.249	...	0	4	2	1	0	2	
675	29	19	5	4	3	225	15	15	41	237.656	...	0	4	2	1	0	2	
681	29	14	5	5	3	225	15	15	41	237.656	...	0	4	2	1	0	2	
683	29	22	5	6	3	225	15	15	41	237.656	...	0	4	2	1	0	2	

5 rows × 21 columns

```
In [19]: data[data['ID']==28].head()
```

Out[19]:

	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	...	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	W
52	28	23	9	3	4	225	26	9	28	241.476	...	0	1	1	0	0	2	
56	28	18	9	4	4	225	26	9	28	241.476	...	0	1	1	0	0	2	
67	28	23	10	6	4	225	26	9	28	253.465	...	0	1	1	0	0	2	
69	28	23	10	4	4	225	26	9	28	253.465	...	0	1	1	0	0	2	
73	28	23	10	4	4	225	26	9	28	253.465	...	0	1	1	0	0	2	

5 rows × 21 columns

```
In [21]: # changing row 51 ID
         data.iloc[51,0] = 28
```

On further inspection, we noticed that these columns data actually matched with ID 28's data, indicating a possible data entry error in the ID field.

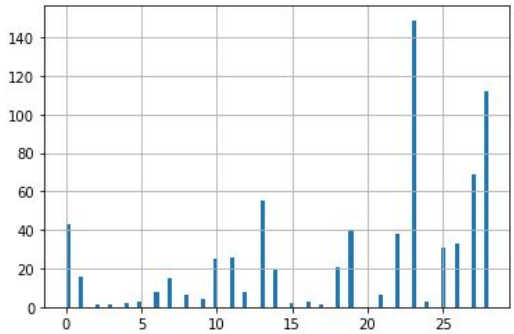
So we correct this error for subsequent analysis by changing ID to 28 for row 51.

2. Medical Consultations Dominate, but not in hours

Most of the instances of absenteeism are caused due to medical, dental consultations, follow up and therapy (22,23,27,28). However they take up very less hours on average. So medical consultations are not a major factor in the absenteeism. Musculoskeletal has the highest total and number of hours product.

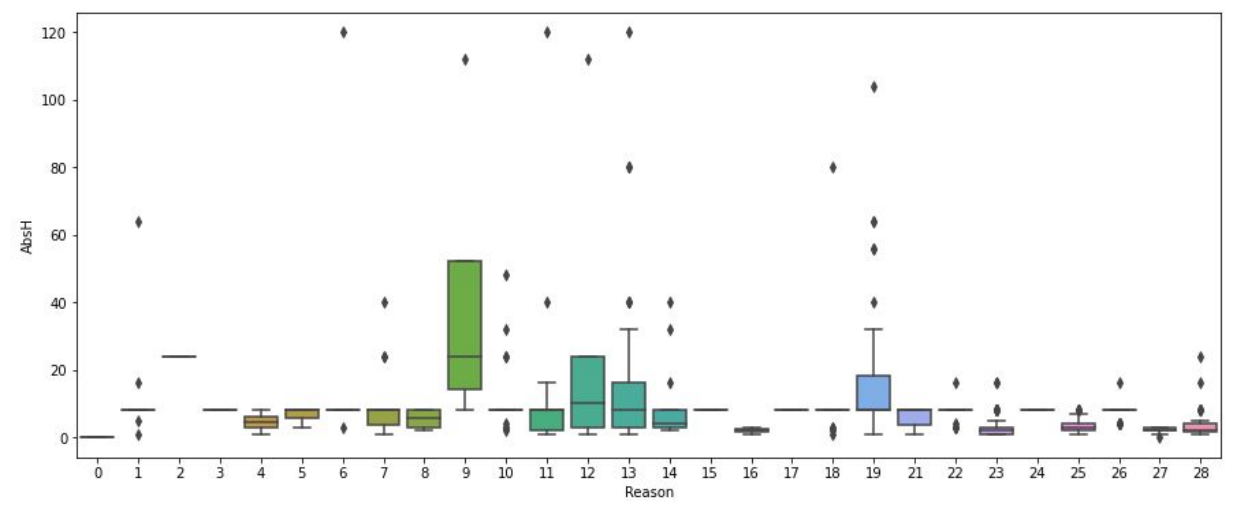
```
In [97]: data.groupby(['Reason'])['AbsH'].mean()
Out[97]: Reason
         0      0.000000
         1     11.375000
         2     24.000000
         3      8.000000
         4      4.500000
         5      6.333333
         6     21.375000
         7     10.000000
         8      5.333333
         9     42.000000
         10    11.040000
         11    11.423077
         12    23.375000
         13    15.309091
         14     8.789474
         15     8.000000
         16     2.000000
         17     8.000000
         18    10.333333
         19    18.225000
         21     5.833333
         22     7.710526
         23     2.845638
         24     8.000000
         25     3.483871
         26     7.272727
         27     2.275362
         28     2.991071
```

```
In [30]: plt.figure()
         data['Reason'].hist(bins = 100)
         plt.show()
```



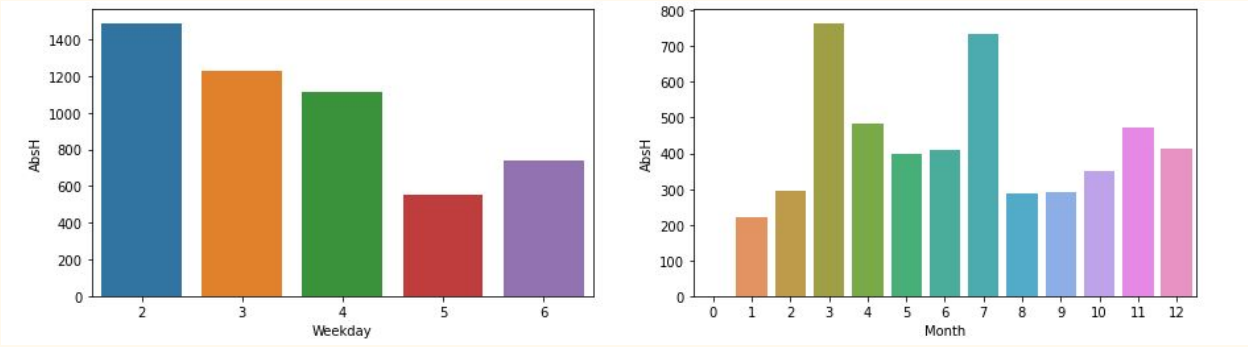
3. Skin, Circulatory diseases cause more absent hours

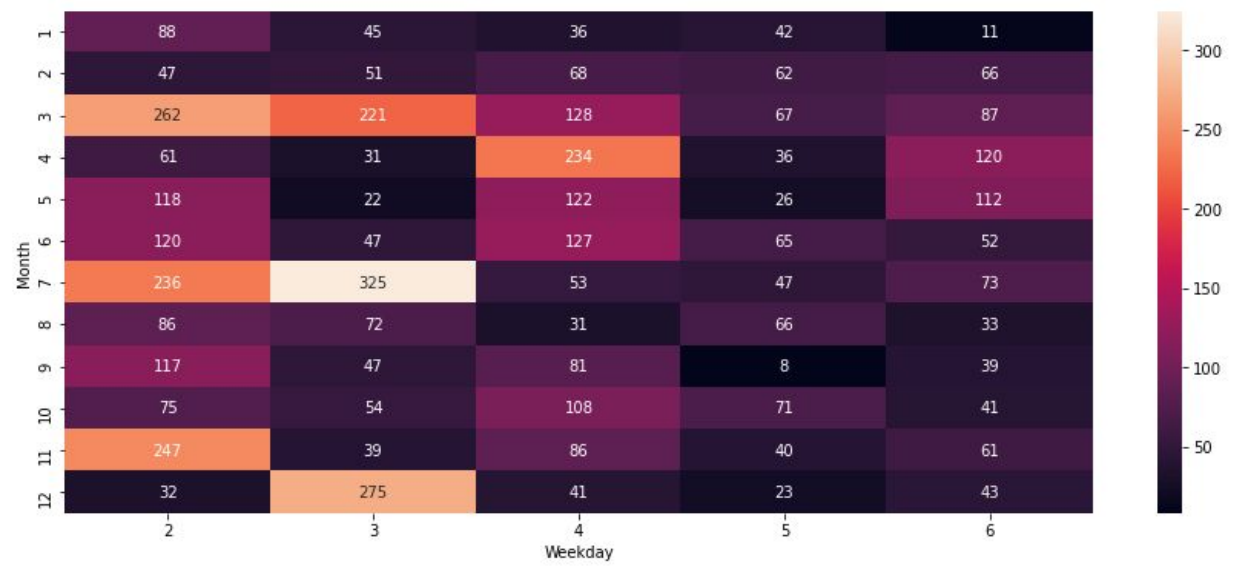
People suffering from these diseases had low instances of absenteeism, but each instance took a large number of hours. This company should do medical checkups in this area, since employees who generally aren't absent are taken off for longer due to these diseases.



4. Monday is the most frequent absent day, while March and July are the months in the same respect.

We find most of the absenteeism hours taking place on Mondays, across all months. The prospect of extending the weekend in the lack of 'Monday motivation' signals that the company needs to improve its working conditions so that people become eager to work fruitfully on all working days.





March is the month with the Holy Week and Easter, a religious occasion for Christians, accounting for 86.8% of the population. After some research we found that some of the Holy Week days are declared holiday, but people might want to take leave during other parts of that week, or extend after Easter.

July is in the middle of Winter in Brazil, which is generally marked by more leaves due to inclement weather and ailments.

5. Multiple people take leave during certain days or months, reducing workforce capacity

Most employees are absent on Wednesdays in May, followed by Mondays and Tuesdays in July and Tuesdays in September. This knowledge can help the company if they are able to track events in the past that took place on those days, and hire extra staff on contract to compensate for the reduced capacity.

```
In [125]: plt.figure(figsize = (15, 6))
          sns.heatmap(data.groupby(['Month', 'Weekday'])['ID'].nunique().unstack()[1:13], annot = True, fmt = 'g')
          plt.show()
```

6. Obese employees have longer absent periods

We find a increasing progression in mean absentee hours when going from normal to overweight to obese. This gives a clear signal for the company to proactively take up measures for exercise, since courier jobs apart from the delivery are desk jobs where people sit for a huge portion of time.

```
In [134]: bin2 = [19,24,29,38]
          labels2 = ['Normal','Overweight','Obese']
          data['BMI_fact'] = pd.cut(data['BMI'],bins=bin2,labels=labels2)
          BMI_sum = data.groupby(['ID','BMI_fact'],as_index=False)['AbsH'].sum()

          BMI_sum.isnull().sum()

Out[134]: ID          0
          BMI_fact    0
          AbsH       72
          dtype: int64

In [135]: BMI_sum.dropna(inplace = True)

In [137]: BMI_hours_missed=round(BMI_sum.groupby('BMI_fact')[['AbsH']].mean(),2)
          BMI_hours_missed
```

Out[137]:

BMI_fact	AbsH
Normal	113.77
Overweight	137.93
Obese	162.33

7. Employees 4, 8, 35 have not taken any leaves

These employees have zero absentee hours. Another correlation spotted is when the reason is 0, we find that disciplinary failure is 1.

```
In [122]: hours_absent = ID_Group.sum()['Absenteeism time in hours']
          hours_absent

Out[122]: ID
          1     121
          2      25
          3     482
          4       0
          5     104
          6      72
          7      30
          8       0
          9     262
          10    186
          11    450
          12     34
          13    183
          14    476
          15    253
          16     16
          17    126
          18    118
          19      6
          20    306
          21     16
          22    253
          23     40
          24    254
          25     42
          26     83
          27     27
          28    347
          29     21
          30     31
          31     16
          32     16
          33     73
          34    344
          35      0
          36    311
```

8. Farther commute distance leads to higher injury and poisoning

There are 29 instances of people taking leave for injury and poisoning where the distance from home is more than 25 kms, as opposed to 11 instances where people took leave for this reason when their travel from their home to work was less than 25kms.

This suggests that farther commute distances cause more injuries due to longer time spent in high traffic roads of Brazil. A possible explanation for poisoning could be that beyond 25 km, the outskirts of the city may not have proper sanitation or hygiene conditions.

```
data[data['Reason'] == 19].groupby(['Distance']).size()
```

```
Distance
10    4
11    1
12    2
13    2
15    1
20    1
25    4
26    4
27    2
29    1
36    8
49    1
50    4
51    3
52    2
dtype: int64
```

9. People in their 30s and 50s are absent more, while the 40s are most regular

In [127]:
```
bins = [20,29,39,49,59]
labels=['Adult20s','Adult30s','Adult40s','Adult50s']
data['age_fact']=pd.cut(data['Age'],bins=bins,labels=labels)
```

In [128]:
```
hours_sum = data.groupby(['ID','age_fact'],as_index=False)['AbsH'].sum()
hours_sum.isnull().sum()
```

Out[128]:
```
ID            0
age_fact      0
AbsH        107
dtype: int64
```

In [129]:
```
hours_sum.dropna(inplace = True)
```

In [130]:
```
age_hours_missed=round(hours_sum.groupby(['age_fact'])[['AbsH']].mean(),2)
age_hours_missed
```

Out[130]:

age_fact	AbsH
Adult20s	118.17
Adult30s	178.44
Adult40s	88.27
Adult50s	147.25

Thank you