



Employee Absenteeism



Smit Patel & Lauren Foltz



Data and Preprocessing

- ❖ Dataset was obtained from the UCI Machine Learning Repository. It contains information regarding a number of employees that were absent between July 2007- July 2010 at a courier company in Brazil.
- ❖ It consisted of an Excel file with 740 rows and 21 columns.
- ❖ Each row is an absence of one of 33 employees.



Data and Preprocessing

- ★ After reading in the data, the dimensions were checked and the output was as expected (740, 21).
- ★ Missing values on the dataset was checked using the function “missing_values_table” developed by Nikos Tavoularis on [StackOverflow.com](https://stackoverflow.com).
- ★ No missing values were found in the dataset.

```
missing_values_table(project)
```


```
Your selected dataframe has 21 columns.  
There are 0 columns that have missing values.
```

Missing Values	% of Total Values
----------------	-------------------

Data and Preprocessing

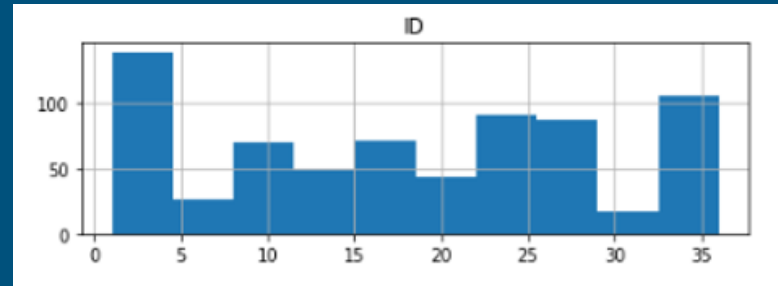
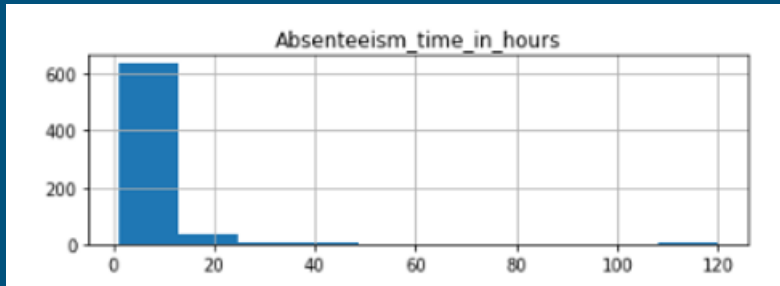
- ❖ A subset was created with just the 11 columns of interest.
- ❖ 44 Rows showed 0 hours missed, but there was a month, day, and season listed. This data appeared dirty and was removed.
- ❖ After preprocessing dataset size was reduced to 696 observations and 11 variables. The dimensions were checked to confirm this, and the output was as expected (696, 11)

ID	Month of absence	Day of the week	Seasons	Service time	Age	Son	Social drinker	Social smoker	Body mass index	Absenteeism time in hours
13	9	4	4	12	31	3	1	0	25	0
33	9	6	4	14	47	2	0	0	32	1
3	9	6	4	18	38	0	1	0	31	3
20	9	6	4	11	36	4	1	0	23	4



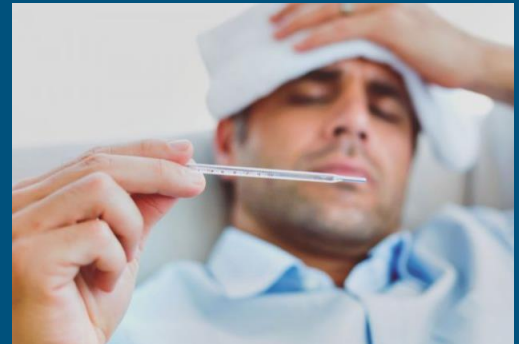
Data and Preprocessing

- ❖ Histograms were created for the 11 columns of interest.
- ❖ They show that most absences were for a smaller number of hours.
- ❖ They also reveal that some employees had many more absences than others.



Research Questions

1. Which Age Group Misses the Most Hours?
2. Which BMI Group Misses the Most Hours?
3. Do Smokers or Nonsmokers Miss More Hours?
4. Do Drinkers or Non-drinkers Miss More Hours?
5. Do Parents or Non-parents Miss More Hours?
6. Which day of the week is missed most?
Which month? Which season?

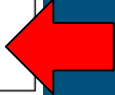


1. Which Age Group Misses the Most Hours?

```
bins = [20,29,39,49,59] #Creating bins
labels=['Adult20s','Adult30s','Adult40s','Adult50s'] #Labelling bins
new_project1['age_fact']=pd.cut(new_project1['Age'],bins=bins,labels=labels) #Creating new column with bins
hours_sum=new_project1.groupby(('ID','age_fact'),as_index=False)['Absenteeism_time_in_hours'].sum() #Sum number of hours
hours_sum=hours_sum.dropna() #Removing NA's that are generated during the process
age_hours_missed=round(hours_sum.groupby('age_fact')[['Absenteeism_time_in_hours']].mean(),2) #Calculating AVG hours
age_hours_missed #Avg hours missed by employees based on Unique ID.
```

Age	Age Bin
20-29	Adult20s
30-39	Adult30s
40-49	Adult40s
50-59	Adult50s

Absenteeism_time_in_hours	
age_fact	
Adult20s	141.80
Adult30s	190.33
Adult40s	97.10
Adult50s	196.33

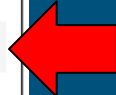


2. Which BMI Group Misses the Most Hours?

```
bin2=[19,24,29,38]           #Binning BMI
labels2=['Normal','Overweight','Obese'] #Labelling BMI's
new_project1['BMI_fact']=pd.cut(new_project1['BMI'],bins=bin2,labels=labels2) #Creating new column to represent BMI value
BMI_sum=new_project1.groupby(('ID','BMI_fact'),as_index=False)['Absenteeism_time_in_hours'].sum() #Sum number of hours
BMI_sum=BMI_sum.dropna() #Removing NA's that are generated during the process
BMI_hours_missed=round(BMI_sum.groupby('BMI_fact')[['Absenteeism_time_in_hours']].mean(),2) #Calculating AVG hours missed
BMI_hours_missed #Avg hours missed by employees based on Unique ID.
```

BMI	BMI Bin
19-24	Normal
25-29	Overweight
30-38	Obese

Absenteeism_time_in_hours	
BMI_fact	
Normal	123.25
Overweight	148.54
Obese	208.71



3. Do Smokers or Nonsmokers Miss More Hours?

```
new_project1['Smoker_cat']=pd.cut(new_project1.Social_smoker,2,labels=['Smoker','Non-Smoker']) #Creating new column
Smoker_sum=new_project1.groupby(('ID','Smoker_cat'),as_index=False)['Absenteeism_time_in_hours'].sum() #Sum number of hours
Smoker_sum=Smoker_sum.dropna() #Removing NA's that are generated during the process
Smoker_hours_missed=round(Smoker_sum.groupby('Smoker_cat')[['Absenteeism_time_in_hours']].mean(),2) #Calculating AVG hours
Smoker_hours_missed #Avg hours missed by employees based on Unique ID.
```


Absenteeism_time_in_hours	
Smoker_cat	
Smoker	183.58
Non-Smoker	50.14



4. Do Drinkers or Non-drinkers Miss More Hours?

```
new_project1['Drinker_cat']=pd.cut(new_project1.Social_drinker,2,labels=['Drinker','Non-Drinker']) #Creating new column
Drinker_sum=new_project1.groupby(('ID','Drinker_cat'),as_index=False)['Absenteeism_time_in_hours'].sum() #Sum
Drinker_sum=Drinker_sum.dropna() #Removing NA's that are generated during the process
Drinker_hours_missed=round(Drinker_sum.groupby('Drinker_cat')[['Absenteeism_time_in_hours']].mean(),2) #Calculating AVG
Drinker hours missed #Avg hours missed by employees based on Unique ID.
```

Absenteeism_time_in_hours	
Drinker_cat	
Drinker	118.62
Non-Drinker	189.76




5. Do Parents or Non-parents Miss More Hours?

```
bin3=[-np.inf,0,np.inf] #Creting bins for 0-4 Levels in "son" column
new_project1['son_fact']=pd.cut(new_project1.Son,bins=bin3,labels=['None','some']) #Creating new column and assigning Lables
Son_sum=new_project1.groupby(('ID','son_fact'),as_index=False)['Absenteeism_time_in_hours'].sum() #Sum number of hours
Son_sum=Son_sum.dropna() #Removing NA's that are generated durig the process
Son_hours_missed=round(Son_sum.groupby('son_fact')[['Absenteeism_time_in_hours']].mean(),2) #Calculating AVG hours missed
Son_hours_missed
```

Number of Children	Parent Bin
0	Non-Parent
>0	Parent

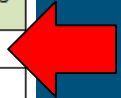
Absenteeism_time_in_hours	
son_fact	
None	133.75
some	167.57



6A. Day of the Week

Keep in mind that if a person called out for 40 hours, the data set only shows the day of week for the start of the absence.

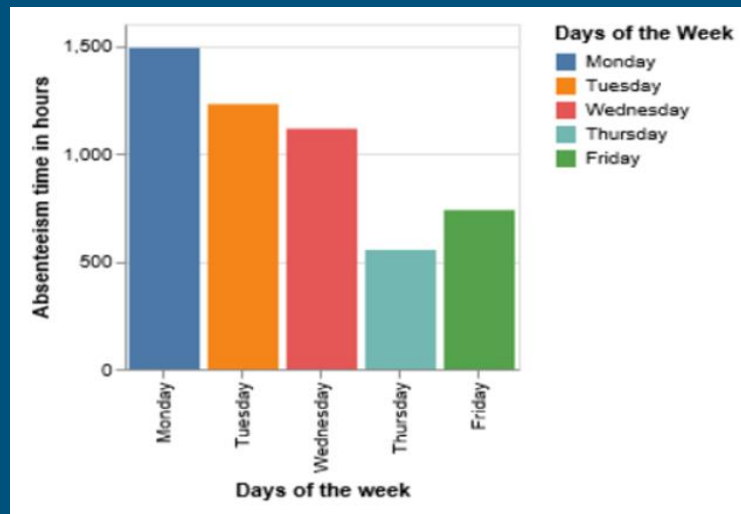
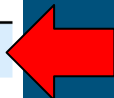
ID	Month of absence	Day of the week	Seasons	Service time	Age	Son	Social drinker	Social smoker	Body mass index	Absenteeism time in hours
14	7	2	1	14	34	2	1	0	25	40
1	7	2	1	14	37	1	0	0	29	8
20	7	2	1	11	36	4	1	0	23	8
20	7	3	1	11	36	4	1	0	23	8



6A. Day of the Week

```
alt.Chart(new_project4).mark_bar().encode(
  alt.X('Day_Factor:N',axis=alt.Axis(title='Days of the week'),sort=['Monday','Tuesday','Wednesday','Thursday','Friday']),
  alt.Y('sum(Absenteeism_time_in_hours):Q',axis=alt.Axis(title='Absenteeism time in hours')),
  color=alt.Color('Day_Factor:N',title='Days of the Week',sort=['Monday','Tuesday','Wednesday','Thursday','Friday'])
).properties(width=200,height=200)
```

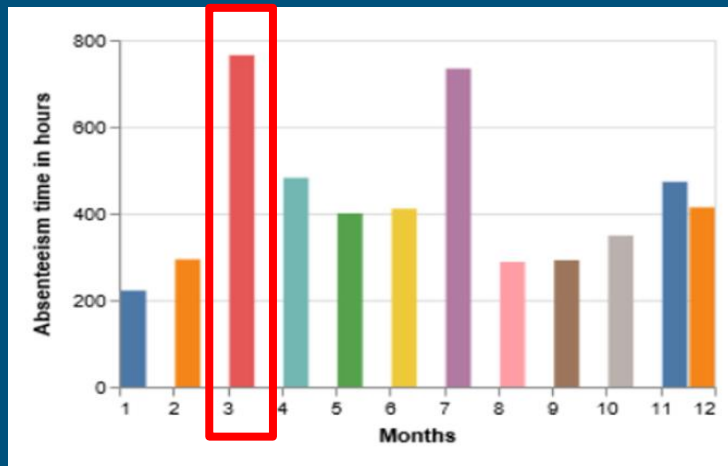
Absenteeism_time_in_hours	
Day_Factor	
Monday	1489
Tuesday	1229
Wednesday	1115
Friday	738
Thursday	553



6B. Month

```
alt.Chart(new_project1).mark_bar().encode(  
    alt.X('Month_of_absence:Q',axis=alt.Axis(title='Months',ticks=True),bin=alt.Bin(maxbins=30)),  
    alt.Y('sum(Absenteeism_time_in_hours):Q',axis=alt.Axis(title='Absenteeism time in hours')),  
    color=alt.Color('Month_of_absence:N',legend=None)  
) .properties(width=300,height=200)
```

March and July missed by most employees:

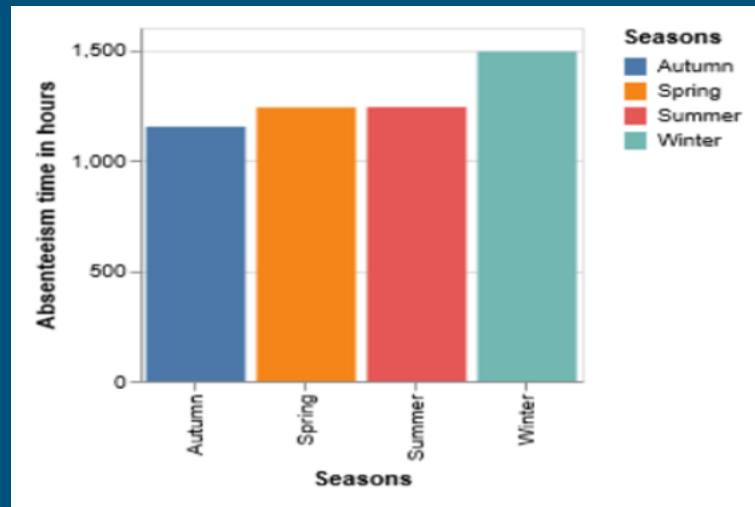
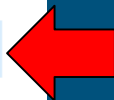


6C. Season

```
alt.Chart(new_project4).mark_bar().encode(  
  alt.X('seasons_fact:N',axis=alt.Axis(title='Seasons')),  
  alt.Y('sum(Absenteeism_time_in_hours):Q',axis=alt.Axis(title='Absenteeism time in hours')),  
  color=alt.Color('seasons_fact:N',title='Seasons')  
) .properties(width=200,height=200)
```

#Selecting Bar chart
#Assigning data to x-axis
#Assigning data to Y-axis
#Assigning color
#Assigning height and width

Absenteeism_time_in_hours	
seasons_fact	
Summer	1241
Autumn	1152
Winter	1492
Spring	1239



Conclusions

1. Adults in their 50s missed the most work.
 2. Obese workers missed the most work.
 3. Smokers missed more work than non-smokers.
 4. Non-drinkers missed more work than drinkers.
 5. Parents missed more work than non-parents.
 6. Workers missed work starting on Monday the most often.
- The highest months for missed work were March and July.
The highest season for missed work was Winter.

Recommendations

- Obesity is associated with more missed days. Encourage workers to maintain a healthy body mass index. Offer incentives or support, such as a weight watchers club on site.
- Be aware that older workers and parents may miss more work, and plan for more substitute coverage if the workforce starts to skew heavily in these areas.
- Be aware that missed work most often starts on Mondays, and more work is missed in March and July. Plan for more substitute coverage at these times. Encourage workers to schedule non-emergency doctor visits towards the end of the week, when fewer days are missed.

Going Forward

- It would be nice if we had the ages of the children. That way we could investigate whether parents of preschool age children miss more work than parents of older children.
- It would be useful to have data on the actual days of the week missed, not just the starting day. This could be helpful for planning coverage.
- Although this data set covered 740 instances over 3 years, it only covered 33 employees. It would be good if we could obtain a larger data set, to see if the results are consistent. We could also split a larger set into training and testing sets, and try to predict who was likely to miss more work, and what the most important variables are for the prediction.
- Which reasons caused the most missed hours?
- If the same trend of absenteeism continues, how much loss of hours can we project for next year?

Going Forward

- We will add our code to Kaggle/Notebook after presentation. Below will be the link to check our work and also track future progress of the project:

<https://www.kaggle.com/allinolbazinga/employee-absenteeism>

